

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

Навчально-науковий інститут телекомунікаційних систем

Кафедра телекомунікацій

До захисту допущено:

Завідувач кафедри

_____ Сергій КРАВЧУК

«__» _____ 2024 р.

Дипломна робота

на здобуття ступеня бакалавра

**за освітньо-професійною програмою «Інженерія та програмування
інфокомунікацій»**

спеціальності 172 Телекомунікації та радіотехніка

**на тему: «Аналіз інструментів для валідації даних для програмного
забезпечення в сфері телекомунікацій»**

Виконала:

студентка IV курсу, групи ТЗ-01

Зінькевич Богдана Романівна _____

Керівник:

науковий керівник НН ІТС, професор кафедри ТК НН ІТС,

д.т.н., професор, академік НАН України

Ільченко Михайло Юхимович _____

Консультант:

Асистент кафедри ТК НН ІТС

Чуб Михайло Миколайович _____

Рецензент:

Доцент кафедри ІТТ НН ІТС

Правило Валерій Володимирович, к.т.н., доцент _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.

Студентка _____

Київ – 2024 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут телекомунікаційних систем
Кафедра телекомунікацій

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 172 «Телекомунікації та радіотехніка»

Освітньо-професійна програма «Інженерія та програмування інфокомунікацій»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Сергій КРАВЧУК

«___» _____ 2024 р.

ЗАВДАННЯ

на дипломну роботу студентці

Зінькевич Богдані Романівні

1. Тема роботи «Аналіз інструментів для валідації даних для програмного забезпечення в сфері телекомунікацій», керівник роботи Ільченко Михайло Юхимович, д.т.н., професор, академік НАН України, затверджені наказом по університету від «22» травня 2024 р. № 2064-с.
2. Термін подання студенткою роботи 10 червня 2024 р.
3. Вихідні дані до роботи: інструменти для валідації якості даних, набори даних
4. Зміст роботи:
 1. Огляд теоретичних аспектів валідації даних
 2. Аналіз інструментів для валідації даних
 3. Тестування роботи вибраних інструментів
 4. Результати роботи інструментів
5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо)

Слайд №1 Вступ: тема роботи, актуальність проблеми, мета і завдання дослідження, об'єкт і предмет дослідження

Слайд №2 Теоретичні аспекти валідації даних

Слайд №3 Огляд популярних інструментів

Слайд №4 Методика тестування

Слайд №5 Критерії оцінки інструментів

Слайд №6 Створення датасетів

Слайд №7 Перевірка якості даних вибраними інструментами

Слайд №8 Оцінка роботи інструментів зі згенерованими даними

Слайд №9 Висновки з експериментального аналізу

Слайд №10 Загальні висновки по роботі

6. Консультанти розділів роботи*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Розділ №1	Консультант Асистент Чуб М.М.	10.01.2024	15.02.2024
Розділ №2	Консультант Асистент Чуб М.М.	15.02.2024	15.03.2024
Розділ №3	Консультант Асистент Чуб М.М.	15.03.2024	10.04.2024
Розділ №4	Консультант Асистент Чуб М.М.	10.04.2024	01.06.2024

7. Дата видачі завдання 10.01.2024 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1.	Визначення мети та постановка завдань дослідження.	10.01.2024 - 16.01.2024	Виконано
2.	Аналіз проблематики та визначення основних положень роботи.	17.01.2024 - 23.01.2024	Виконано
3.	Аналіз доступних рішень для валідації даних.	24.01.2024 - 30.01.2024	Виконано
4.	Розділ 1: Теоретичні аспекти валідації даних.	31.01.2024 - 15.02.2024	Виконано
5.	Розділ 2: Аналіз інструментів для валідації даних.	16.02.2024 - 15.03.2024	Виконано
6.	Розділ 3: Тестування роботи вибраних інструментів.	16.03.2024 - 10.04.2024	Виконано
7.	Розділ 4: Результати роботи інструментів.	11.04.2024 - 01.06.2024	Виконано
8.	Загальні висновки по роботі.	02.06.2024 - 05.06.2024	Виконано
9.	Оформлення дипломної роботи.	06.06.2024 - 10.06.2024	Виконано
10.	Підготовка звітних та презентаційних матеріалів.	11.06.2024 - 16.06.2024	Виконано

Студентка

Богдана ЗІНЬКЕВИЧ

Керівник

Михайло ІЛЬЧЕНКО

* Якщо визначені консультанти. Консультантом не може бути зазначено керівника дипломної роботи.

РЕФЕРАТ

Робота містить 128 сторінок, 124 ілюстрації. Для написання було використано 11 джерел.

Метою роботи є дослідження та порівняння популярних інструментів для валідації даних у сфері телекомунікацій.

При виконанні даної роботи було проведено ознайомлення із популярними інструментами на ринку, вибрано та досліджено три з них, також були створені датасети та практично застосовано стандарт ISO 25012.

В результаті даної роботи було повністю охарактеризовано і описано три протестованих інструменти: Data Ladder, Alteryx та Informatica Data Quality, складена порівняльна таблиця та зроблені висновки щодо функціональних можливостей кожного з них. Також були надані інструкції та рекомендації щодо користування даними інструментами.

Перелік ключових слів: валідація даних, інструменти для валідації даних, стандарт якості, критерії оцінки, набір даних.

ABSTRACT

The bachelor's thesis contains 128 pages, 124 illustrations. The bachelor's thesis was written using 11 sources.

The purpose of this paper is to study and compare popular tools for data validation in the telecommunications sector.

In the course of this work, we reviewed the popular tools on the market, selected and studied three of them, created datasets, and practically applied the ISO 25012 standard.

As a result of this work, the three tested tools were fully characterized and described: Data Ladder, Alteryx, and Informatica Data Quality, and a comparative table was compiled and conclusions were drawn regarding the functionality of each of them. Instructions and recommendations on how to use these tools were also provided.

Keywords: data validation, tools, quality standard, evaluation criteria, data set.

ЗМІСТ

ВСТУП.....	8
РОЗДІЛ 1	13
ТЕОРЕТИЧНІ АСПЕКТИ ВАЛІДАЦІЇ ДАНИХ.....	13
1.1 Визначення основних понять і термінів	13
1.2 Проблематика валідації даних	15
1.3 Стандарт якості даних ISO 25012	19
Висновки	25
РОЗДІЛ 2	26
АНАЛІЗ ІНСТРУМЕНТІВ ДЛЯ ВАЛІДАЦІЇ ДАНИХ	26
2.1 Огляд популярних інструментів для валідації даних на ринку	26
2.2 Методика тестування	30
2.3 Критерії оцінки інструментів валідації даних	32
Висновки	38
РОЗДІЛ 3	40
ТЕСТУВАННЯ РОБОТИ ВИБРАНИХ ІНСТРУМЕНТІВ.....	40
3.1 Створення датасетів.....	40
3.2 Перевірка якості даних вибраними інструментами.....	43
Висновки	116
РОЗДІЛ 4	117
РЕЗУЛЬТАТИ РОБОТИ ІНСТРУМЕНТІВ.....	117
4.1 Оцінка роботи інструментів зі згенерованими даними.....	117
4.2 Висновки з експериментального аналізу	128
ЗАГАЛЬНІ ВИСНОВКИ ПО РОБОТІ.....	130
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:	132

ПЕРЕЛІК СКОРОЧЕНЬ

API	Application Programming Interface
CI/CD	Continuous integration/continuous delivery
CRM	Customer relationship management
CSV	Comma-Separated Values
DDT	Data-driven testing
ETL	Extract, Transform, Load
GPT	Generative Pre-trained Transformer
GTP	GPRS Tunneling Protocol
ID	Identity document
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
ODBC	Open DataBase Connectivity
SDK	Software Development Kit
SIP	Session Initiation Protocol
SQL	Structured Query Language
XML	EXtensible Markup Language
XSS	Cross Site Scripting
БД	База даних
СУБД	Система управління базами даних

ВСТУП

Актуальність проблеми

На сьогоднішній день обсяги даних швидко зростають, а програмне забезпечення стрімко розвивається і стає більш складним і так як жодне програмне забезпечення не може існувати окремо від даних, які користувачі можуть вводити, отримувати і зберігати, комбінація некоректних даних може призвести до помилок, затримок і в результаті неправильного функціонування програми, тому питання якості та достовірності даних набувають критичного значення.

Щодня телекомунікаційні компанії працюють з величезними обсягами інформації з різних джерел: статистика, абонентські дані, фінансові транзакції, технічні показники тощо. Однією з ключових причин актуальності цієї проблеми є те, що обсяги та швидкість надходження даних постійно зростають. Телекомунікаційні компанії обробляють мільярди записів щодня, ці дані повинні бути швидко і точно валідовані, щоб забезпечити їх надійність та придатність для подальшої обробки та аналізу. Різноманіття джерел даних, таких як базові станції, роутери, білінгові системи, додає складності процесу валідації, оскільки кожне джерело має свої особливості форматів даних. Ризики та наслідки помилок в обробці цих даних також підкреслюють важливість даної проблеми. Будь-які помилки, пов'язані з цими даними можуть призвести до серйозних наслідків, включаючи фінансові втрати, зниження якості обслуговування клієнтів та погіршення репутації компанії. Валідація даних у цьому контексті стає однією з найважливіших складових процесу обробки інформації. Крім того, умови жорсткої конкуренції на ринку вимагають від компаній забезпечення максимальної точності та якості даних для прийняття стратегічно важливих рішень, таких як впровадження нових послуг, оптимізація мережі та покращення обслуговування клієнтів.

Також в умовах сучасного розвитку інформаційних технологій дуже затребуваними стали ETL проєкти (Extract - Transform - Load), вони особливо важливі у сферах, де потрібно обробляти великі обсяги даних, взяті з різних джерел і перетворювати їх для подальшого зберігання та аналізу. Основним завданням ETL систем є структурування, збагачення, оптимізація та передача вихідних даних з кількох програмних оболонок у єдину централізовану базу для їх зберігання та застосування.[1] Завдяки тому, що інформація наводиться в єдиний формат вона стає придатною для подальшої обробки. Вкрай важливо відслідковувати, щоб під час трансформації дані не втрачали своєї якості і не спотворювались. Валідація як вихідних, так і оброблених даних є обов'язковою складовою у всіх сферах. Важливо, щоб обраний інструмент для трансформації даних не лише ефективно виконував свої функції, але й забезпечував збереження якості даних.

Ретельний підбір та впровадження ефективних інструментів валідації сприяють підвищенню надійності та точності даних, що є запорукою успішної діяльності телекомунікаційних компаній у сучасних умовах. Аналіз сучасних інструментів валідації даних дозволить визначити найкращі методи і технології для забезпечення високої якості даних, що, у свою чергу, сприятиме підвищенню ефективності та конкурентоспроможності телекомунікаційних компаній.[1]

Мета і завдання дослідження

Метою даної роботи є ґрунтовний аналіз та порівняння популярних інструментів для валідації даних, які можуть бути використані у телекомунікаційній галузі. Для досягнення цієї мети необхідно визначити основні показники та параметри за якими будуть оцінені вибрані інструменти, перевірити чи відповідають вони вимогам стандарту ISO 25012, оцінити кожен інструмент за визначеними показниками якості в роботі із згенерованими даними задля більшого вивчення ефективності інструментів у різних сценаріях.

Потрібно порівняти функціональну придатність цих інструментів до конкретних завдань у сфері телекомунікацій. Ключові аспекти, які необхідно розглянути, це налаштування правил оцінки, робота з обсягами даних, обробка різноманітних форматів даних, моніторинг якості даних у режимі реального часу та відповідність нормативним вимогам.

В даній роботі ми ознайомимось з такими сценаріями тестування даних, як перевірка цілісності даних, тестування введення та виведення даних, збереження, обробка, відновлення, а також реплікація даних. Все це ми зможемо здійснити завдяки таким методам тестування даних, як ETL тестування, процесу перевірки правильності вилучення, перетворення та завантаження даних між різними системами. Data Quality тестування, тобто оцінимо якість даних з точки зору їхньої точності, повноти, узгодженості, достовірності та актуальності. Також згенеруємо тестові набори даних, які зможуть бути використані для тестування різних сценаріїв і різних інструментів. Всі ці кроки спрямовані на досягнення ще однієї мети, це більш детальне ознайомлення і застосування стандарту ISO 25012. Це стандарт, що приносить значну користь для організацій і розробників і часто стає “точкою відліку для вашої роботи”.[2]

Остання, але не менш важлива мета, це ознайомити спільноту з інструментами тестування даних та сценаріями самого тестування, яка є важливою частиною складової процесу розробки програмного забезпечення. Для цього ми звернемо увагу на такі аспекти, як важливість тестування даних у циклі розробки ПЗ, огляд сценаріїв тестування даних та рекомендації щодо вибору та впровадження інструментів залежно від конкретних умов проекту.

Очікується, що дане дослідження зможе забезпечити більш детальне розуміння можливостей і обмежень, переваг і недоліків різних інструментів для валідації даних, їх відповідності вимогам сфери телекомунікаційної галузі та стандарту ISO 25012.

Об'єкт і предмет дослідження

Об'єктом даного дослідження є процес забезпечення якості та цілісності даних в інформаційних системах, який охоплює весь життєвий цикл даних, починаючи від їх введення, оброблення, передачі, зберігання та завершуючи видаленням або архівацією. До основних аспектів процесу забезпечення якості та цілісності даних належать підтримка коректності та достовірності відповідно до встановлених правил та обмежень для різних типів даних. Забезпечення повноти даних, уникнення втрат та дублікатів, дотримання цілісності даних, підтримка актуальності та своєчасного оновлення даних, а також захист від несанкціонованого доступу, модифікації або знищення і можливість відновлення даних у разі збоїв або пошкоджень. Це багатогранний процес, що вимагає ретельного підходу та застосування різноманітних інструментів і методів для підтримки коректності даних.

Тому предметом даного дослідження є інструменти та методи валідації даних, які є ключовими для забезпечення якості та цілісності даних в інформаційних системах. Кожен інструмент буде досліджений в межах його функціональних можливостей, тобто оцінимо спектр функцій, які підтримують інструменти, аналіз гнучкості налаштувань та можливостей адаптації до специфічних вимог. Розглянемо продуктивність кожного з них, а саме швидкість обробки даних, а також аналіз ефективності ресурсів, необхідних для роботи. Принципи роботи, аналіз надійності та стабільності роботи у відповідності до стандарту. Проводячи тестування з використанням згенерованих даних для додаткових тестів та аналізу поведінки інструментів в різних умовах, включаючи і складні сценарії, зможемо визначити основні переваги, недоліки та обмеження застосування компаніями у сфері телекомунікацій. Також значна увага в даному дослідженні приділяється різним методам та підходам до валідації даних, серед них серверна валідація, яка гарантує, що лише коректні та безпечні дані будуть збережені та використані в системі, валідація на рівні бази даних, що включає використання вбудованих механізмів СУБД для

забезпечення цілісності даних, а також періодичні перевірки, корекція помилок, очищення та стандартизація даних.

РОЗДІЛ 1

ТЕОРЕТИЧНІ АСПЕКТИ ВАЛІДАЦІЇ ДАНИХ

1.1 Визначення основних понять і термінів

Валідація даних є критично важливим процесом у телекомунікаційних системах, який забезпечує цілісність, достовірність та безпеку даних, що обробляються та передаються через ці системи.

Валідація даних - це процес перевірки, чи відповідають вхідні дані заздалегідь визначеним критеріям, правилам та обмеженням.

Під час валідації дані перевіряються на відповідність певним вимогам, наприклад, типам даних, діапазонам значень, форматам, довжині та відсутності неприпустимих символів.

Валідація гарантує, що дані, які обробляються та передаються телекомунікаційними системами, є коректними та узгодженими, запобігаючи некоректній поведінці чи збоям. Також належна валідація даних допомагає уникнути помилок та неочікуваної поведінки, забезпечуючи стабільну та надійну роботу телекомунікаційних систем.

Валідація даних є ключовим елементом захисту телекомунікаційних систем від різних видів атак, запобігаючи можливим витокам даних чи компрометації систем. Багато галузевих стандартів та регуляторних вимог у телекомунікаціях вимагають належної перевірки для забезпечення безпеки, конфіденційності та цілісності даних.

Валідація даних це процес, який допомагає виявляти та усувати помилки, підвищуючи якість та достовірність даних для подальшої обробки та аналізу. Крім того, ефективна валідація забезпечує зрозумілі повідомлення, що допомагає користувачам виправляти неправильно введені дані, покращуючи загальний досвід взаємодії з системою.

Валідація включає в себе різні види перевірок вхідних даних для забезпечення їх коректності, цілісності та безпеки.

Перше, це перевірка типів даних, ця перевірка полягає в тому, щоб переконатися, що вхідні дані відповідають очікуваному типу, наприклад, чи є вхідне значення цілим числом, або ж чи є вхідне значення дати або часу у відповідному форматі.

Наступна - перевірка діапазонів значень. Перевірка діапазонів забезпечує відповідність вхідних значень межах допустимого діапазону. Гарним прикладом, який демонструє цю перевірку є віковий діапазон. Якщо уявити, що розробляється інтерфейс програми чи веб сайту, де користувач повинен ввести свій вік у відповідне поле, валідація вікового діапазону дозволяє переконатись, що користувач ввів допустиме значення, яке знаходиться в межах 0-120 років. Така перевірка допомагає уникнути некоректних значень, які можуть вплинути на роботу системи.

Перевірка форматів даних, полягає в тому, щоб переконатись, що вхідні дані відповідають певним шаблонам. Наприклад, електронні адреси та телефонні номери. Під час практичного дослідження я планую звернути особливу увагу на те, як впораються інструменти з даними неправильного формату, а саме електронними поштами та номерами телефонів.

Перевірка на наявність заборонених/небезпечних символів, іншими словами, виключення символів, які можуть бути використані для атак або викликати помилки у системі.

Перевірка довжини даних, забезпечує, що вхідні дані не перевищують встановлену максимальну довжину, наприклад, чи не перевищує довжина текстового поля встановлений максимум у 255 символів, або ж чи не перевищує числове значення допустиму кількість цифр.

Перевірка залежностей між полями, цей вид перевірки забезпечує логічну узгодженість між різними полями даних. Наприклад, чи дата початку не є пізнішою за дату завершення, або ж чи значення одного поля відповідає

вимогам, заснованим на значеннях інших полів. Наприклад, якщо у формі є поля "Тип пасажирів" та "Дата народження". Якщо "Тип пасажирів" = "Дитина", то поле "Дата народження" має відповідати певним критеріям, що визначають дитячий вік.

Остання в цьому списку, але не менш важлива, це перевірка безпеки, що включає в себе заходи для захисту систем від атак, зокрема SQL-ін'єкцій та XSS (міжсайтового скриптингу). SQL-ін'єкція — це тип атаки, при якій зловмисник вставляє або "ін'єктує" шкідливі SQL-запити в поле вводу даних, які потім виконуються на сервері бази даних. Це може призвести до викрадення, видалення даних або навіть повного захоплення контролю над базою даних. Інший вид, це XSS — тип атаки, при якому зловмисник вставляє шкідливі скрипти у веб сторінку, яка потім відображається іншим користувачам. Ці скрипти можуть викрадати інформацію про користувача, зокрема сесійні дані, або виконувати інші небажані дії від імені користувача.

Валідація даних є невід'ємною частиною розробки та експлуатації в телекомунікаційній сфері, забезпечуючи надійність, безпеку та коректну роботу цих критично важливих систем, що обробляють величезні обсяги даних. Без належної валідації можуть виникати помилки в обробці даних, коли некоректні або пошкоджені дані призводять до неправильних результатів обробки, а в тому числі і впливають на якість послуг.[3]

Так як телекомунікаційні системи включають в себе мережеве обладнання, програмне забезпечення для керування мережею, системи білінгу та інші компоненти, вони потребують точних та коректних даних для забезпечення безперебійної роботи.

1.2 Проблематика валідації даних

Проблематика валідації даних включає в себе ряд аспектів і викликів, які можуть виникнути при роботі з обробкою та аналізом інформації. Валідація

даних у сфері телекомунікацій є надзвичайно важливим і водночас складним завданням. Телекомунікаційні системи оперують величезними обсягами різноманітних даних, які надходять з різних джерел та у різних форматах. Крім того, ці системи повинні забезпечувати високу доступність, надійність та безпеку, обробляючи критично важливі дані в режимі реального часу.

Через специфічні вимоги галузі телекомунікацій, процес валідації даних у цій сфері стикається з низкою унікальних проблем та викликів.

Перша проблема, це масштабованість та великі обсяги даних. Телекомунікаційні мережі обробляють колосальні обсяги різноманітних даних, таких як дані про абонентів, деталі викликів, трафік мережі, метадані тощо. Процес валідації таких об'ємів вимагає значних обчислювальних ресурсів і часу, а системи валідації повинні бути масштабованими, щоб впоратися зі зростаючими обсягами даних, не втрачаючи продуктивності.

Наступна проблема, часові обмеження та продуктивність. Валідація великих обсягів даних у режимі реального часу вимагає дуже високої продуктивності від систем валідації. Багато телекомунікаційних даних, таких як дані сигналізації, трафік мережі, оновлення місцезнаходження, повинні оброблятися негайно, без затримок. Необхідно забезпечити швидку обробку даних, мінімізуючи затримки та дотримуючись жорстких часових обмежень. Крім цього для досягнення необхідної продуктивності валідація даних часто вимагає паралельної обробки на кластерах або в розподілених системах, у деяких випадках критичні дані, такі як екстрені виклики, повинні мати пріоритет над іншими даними для забезпечення своєчасної обробки.

Наступною проблемою є різноманітність протоколів та форматів даних. У телекомунікаціях використовується велика кількість різноманітних протоколів, як старих, так і нових, прикладами є SIP, Diameter, GTP, SIGTRAN та багато інших. Кожен протокол має власні специфікації та вимоги щодо форматів даних, кодування та правил валідації. Крім протоколів, дані ще надходять в різних форматах, таких як XLSX, XML, JSON, SQL тощо.

Складність правил валідації є однією з ключових проблем у процесі валідації даних у сфері телекомунікацій. Телекомунікаційна галузь регулюється численними стандартами, нормами та правилами, встановленими регуляторними органами та галузевими організаціями. Крім стандартних правил валідації, телекомунікаційні компанії часто мають специфічні бізнес-правила, які також потрібно враховувати. Ці правила можуть включати перевірку корпоративних політик, тарифних планів, спеціальних пропозицій тощо. Все це додає складності до даного процесу.

Взаємодія з іншими системами є ще однією проблемою.. Телекомунікаційні системи часто інтегровані з іншими системами, такими як білінгові системи, системи управління взаєминами з клієнтами, системи маршрутизації трафіку, системи забезпечення операційної підтримки тощо. Кожна система може мати власні правила валідації даних, формати та протоколи передачі. При обміні даними між різними системами критично важливо забезпечити узгодженість та цілісність цих даних, тому правила валідації повинні бути узгоджені між усіма системами, щоб уникнути конфліктів та некоректної обробки даних, тим паче, що дані можуть передаватися у різних форматах. Процес валідації має враховувати ці трансформації форматів і гарантувати, що дані залишаються коректними після перетворень.

Наступна проблема, це відмовостійкість та резервування. Телекомунікаційні системи повинні забезпечувати високу доступність та безперебійну роботу постійно, обробляючи критично важливі дані різного роду. Будь-які збої або відмови в процесах валідації даних можуть призвести до серйозних наслідків для мережі та клієнтів. Тому системи валідації даних повинні бути стійкими до відмов окремих компонентів або вузлів, забезпечуючи механізми автоматичного відновлення та переходу на резервні компоненти в разі збоїв. Ці резервні компоненти повинні бути готові до роботи в будь-який момент у випадку відмови основних систем.

Ще однією проблемою є обробка помилок та повідомлень. Під час валідації даних часто виникають помилки, такі як невідповідність форматам, порушення діапазонів значень тощо, і так як, у телекомунікаційних системах, які обробляють величезні обсяги даних, може генеруватися величезна кількість повідомлень про помилки, необхідно ефективно керувати цими потоками повідомлень, фільтрувати, категоризувати та визначати їх пріоритетність. Повідомлення про ці помилки повинні бути зрозумілими та інформативними, адже незрозумілі або недостатньо деталізовані повідомлення про помилки можуть ускладнити процес вирішення проблем.

Ресурсоефективність та оптимізація, останній, але не менш важливий аспект при впровадженні систем валідації даних у сфері телекомунікацій. Процеси валідації даних можуть бути ресурсомісткими, споживаючи значну обчислювальну потужність, пам'ять та мережеві ресурси. У телекомунікаційних системах, які оперують величезними обсягами даних, це навантаження на ресурси може бути надзвичайно високим. Для забезпечення оптимальної продуктивності необхідно рівномірно розподіляти навантаження на процеси валідації між доступними ресурсами. Нерівномірний розподіл навантаження може призвести до зниження загальної продуктивності. Крім цього, телекомунікаційні системи повинні бути масштабованими для обробки зростаючих обсягів даних та збільшення кількості користувачів.

Всі перелічені проблеми ілюструють складність та важливість належної валідації даних у телекомунікаційних системах. Вирішення цих проблем вимагає ретельного проектування, використання новітніх технологій, оптимізації процесів та узгодженої роботи різних компонентів телекомунікаційної інфраструктури.

1.3 Стандарт якості даних ISO 25012

Обсяги даних та інформації, що обробляються комп'ютерними системами, зростають у всьому світі. Якість даних є ключовим компонентом якості та корисності інформації, отриманої на основі цих даних, а також більшості бізнес-процесів.

Контроль та підвищення якості даних є надважливим завданням на сьогоднішній день, зокрема через:

- отримання даних від організацій, в яких якість процесу виробництва даних невідома або низька;
- існування дефектних даних, що призводить до незадовільної інформації, непридатних для використання результатів і незадоволених клієнтів;
- розподілення даних між різними власниками та користувачами. Дані, зібрані для потреб однієї організації, часто не мають узгодженого та інтегрованого бачення, необхідного для забезпечення співпраці;
- співіснування застарілої архітектури та комп'ютерних систем з розподіленими системами, розробленими і реалізованими в різний час і за різними стандартами;
- наявність інформаційних систем, де дані часто змінюються, а інтеграція є особливою проблемою.

Модель якості даних, визначена в стандарті ISO 25012, спрямована на задоволення всіх вище перелічених потреб, а також, враховуючи те, що життєвий цикл даних часто довший за життєвий цикл програмного забезпечення вона може бути використана також для:

- визначення та оцінювання вимог до якості даних у процесах виробництва, збору та інтеграції даних
- визначення критеріїв забезпечення якості даних, також корисних для оцінки та покращення даних

- оцінки відповідності даних законодавству та вимогам.

Виявлення помилок або неефективності, пов'язаних з даними, надає можливість для вдосконалення та корекції різних компонентів системи, включаючи:

- дані (наприклад, синтаксичний аналіз, очищення, збагачення, трансформація чи узгодження);
- програмне забезпечення (внесення змін до початкових програм для впровадження контролю узгодженості);
- апаратне забезпечення (модернізація комп'ютерної системи для покращення часу відгуку);
- людські бізнес-процеси (наприклад, навчання користувачів для запобігання помилкам при введенні даних; покращення облікових процесів, які керують даними).

Міжнародний стандарт ISO 25012 встановлює загальну модель якості даних для тих, що зберігаються у структурованому форматі в комп'ютерних системах. Стандарт зосереджений на якості даних як частини комп'ютерної системи та визначає характеристики якості для даних, що використовуються людьми та системами.[4]

На рисунку 1.1 схема загальної структури системи: вона може включати інформаційні системи, які, у свою чергу, можуть включати одну чи декілька комп'ютерних систем.

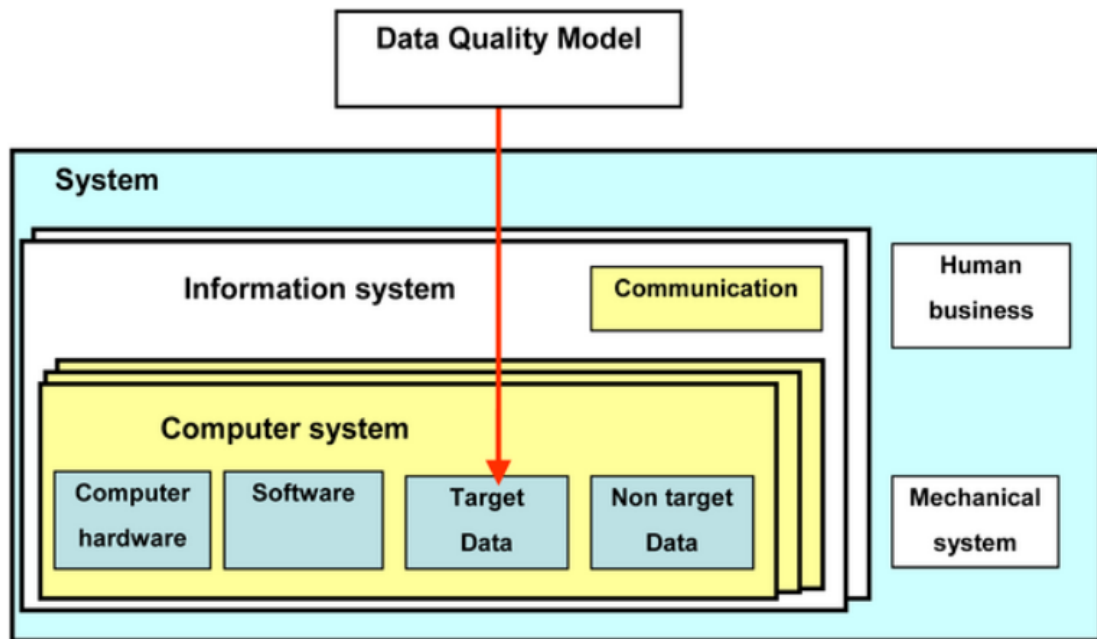


Рис. 1.1 Цільова область моделі якості даних [4]

Міжнародний стандарт ISO 25012 визначає вимоги та показники якості даних, які використовуються для оцінки якості даних незалежно від інших компонентів комп'ютерної системи. Стандарт спрямований на підтримку впровадження процесів життєвого циклу системи і охоплює всі типи даних, включаючи рядки символів, тексти, дати, числа, зображення, звуки тощо, а також їхні значення і зв'язки між ними, наприклад, узгодженість між даними в одній або різних сутностях. Водночас обсяг стандарту не включає дані, створені вбудованими пристроями або датчиками реального часу, які не зберігаються для подальшої обробки або для історичних цілей.

Цей стандарт не регулює фізичну організацію даних, тобто системи керування базами даних. Крім того, проектування концептуальних, логічних та фізичних схем виходить за межі його сфери застосування. Проте, усі процеси та результати, пов'язані з такими даними, отримують користь від застосування цього міжнародного стандарту.[4]

Стандарт якості ISO 25012 визначає характеристики, які допомагають оцінювати та покращувати якість даних (рис. 1.2). Він охоплює як внутрішні

аспекти якості даних, так і залежні від системи характеристики, кожна з яких має важливе значення для забезпечення високої якості даних і повинна враховуватися при проектуванні, впровадженні та експлуатації інформаційних систем.

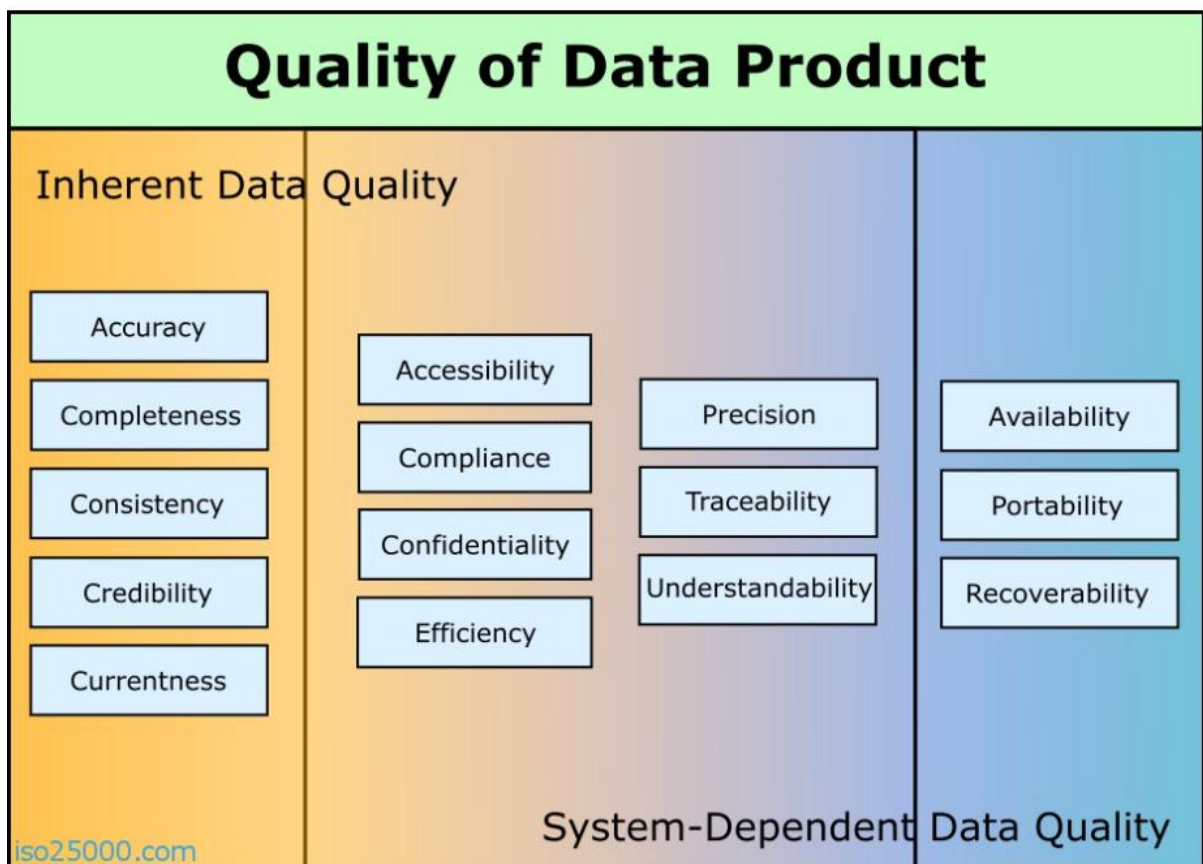


Рис. 1.2 Характеристики стандарту якості ISO 25012 [2]

Існує 3 категорії якості даних, які в сумі складаються з 15 характеристик.

- Власна якість даних (Inherent Data Quality) - визначає, наскільки добре дані відповідають явним і неявним потребам. Складається дана категорія з 5 характеристик:[2]

а. Точність (Accuracy) - відображає наскільки інформація відповідає дійсності в конкретному контексті використання. Має 2 аспекти: синтаксичну та семантичну точність. Синтаксична точність - вимірює наскільки дані відповідають синтаксичним нормам, тобто чи дані

правильної форми. Семантична точність - вимірює, чи правильно дані передають зміст у межах їхньої предметної області.[2]

b. Повнота (Completeness) - визначає наскільки повними є дані стосовно конкретної сутності у визначеному контексті. Дані вважаються повними, якщо вони містять значення для всіх очікуваних атрибутів та екземплярів.

c. Узгодженість (Consistency) - відображає ступінь відповідності даних іншим даним в тому ж контексті використання.

d. Достовірність (Credibility) - відображає ступінь, до якої дані вважаються правдивими та надійними в конкретному контексті використання. Включає концепцію автентичності (правдивість походження, атрибуції, зобов'язань)

e. Актуальність (Currentness) - вказує на ступінь відповідності даних їхній актуальності у визначеному контексті використання.

- Характеристики якості даних, що залежна від системи складають другу категорію і містять всього 3 атрибути: доступність, портативність та відновлюваність. Ці характеристики допомагають забезпечити, щоб дані були доступні користувачам тоді, коли вони потрібні, захищені від несанкціонованого доступу, ефективно використовувалися та легко відновлювалися у випадку втрати або пошкодження.[2]

a. Доступність (Availability) - відображає міру, в якій дані можуть бути отримані авторизованими користувачами.

b. Портативність (Portability) - оцінює здатність даних бути перенесеними в іншу систему без втрати якості.

c. Відновлюваність (Recoverability) - оцінює здатність системи зберігати дані для подальшого використання у випадку помилок або втрат.

- Змішана якість даних (Inherent and System-Dependent Data Quality) — ступінь відповідності даних потребам користувачів та вимогам конкретного програмного забезпечення. Ця категорія охоплює сім атрибутів:

- a. Доступність (Accessibility) - вимірює, наскільки дані доступні в конкретному контексті використання.
- b. Відповідність (Compliance) - оцінює, наскільки дані відповідають встановленим стандартам, домовленостям і правилам.
- c. Конфіденційність (Confidentiality): - визначає, наскільки дані захищені і чи мають атрибути, які забезпечують доступ лише авторизованим користувачам.
- d. Ефективність (Efficiency) - оцінює, наскільки ефективно дані зберігаються та обробляються і скільки часу це займає.
- e. Влучність (Precision) - оцінює можливість чітко розрізняти і розуміти дані.
- f. Відстежуваність (Traceability) - визначає можливість відстежувати доступ до даних і зміни даних.
- g. Зрозумілість (Understandability): - оцінює, наскільки дані можуть бути прочитані та інтерпретовані користувачами, написані відповідними мовами та символами.[2]

Отже, стандарт якості ISO 25012, також відомий як стандарт SQuaRE (Software Product Quality Requirements and Evaluation), є частиною серії стандартів ISO 25000, яка орієнтована на оцінку та управління якістю програмного забезпечення та інших продуктів і систем.[5]

ISO 25012 надає загальний стандарт для оцінки та управління якістю даних, дозволяючи організаціям та розробникам визначати, вимірювати та поліпшувати якість їх даних відповідно до конкретних потреб та контексту використання. Цей стандарт створює основу для систематичного підходу до оцінки якості даних і забезпечує спільний мовний фреймворк для обговорення якісних аспектів даних між зацікавленими сторонами.

Висновки

Валідація даних - це критично важливий процес перевірки точності, повноти, достовірності та відповідності даних встановленим вимогам перед їх використанням в аналізі або прийнятті бізнес-рішень. Методи валідації можуть бути різноманітними і варіюватись від простих перевірок до складних статистичних моделей.

Відсутність належної валідації може призвести до ряду проблем, таких як помилки в аналізі, неправильні висновки та неефективні рішення. Валідація допомагає виявити та виправити помилки на ранніх етапах, мінімізуючи ризики та підвищуючи надійність результатів.

Вагому роль у визначенні якості даних також відіграє ISO 25012, це стандарт, що допомагає оцінювати та покращувати якість даних. Відштовхуючись від стандарту можна протестувати потрібні вам дані і зробити висновок щодо їх якості, або ж покращити неякісні дані. Власне в будь-якому випадку “стандарт — чудова відправна точка” для ваших цілей, пов’язаних з якістю даних.

Отже, ретельна валідація даних є ключовим фактором успіху у сфері телекомунікацій. У цій галузі точність і надійність даних мають особливо критичне значення для забезпечення стабільної роботи мереж, якості обслуговування клієнтів та прийнятті бізнес - рішень.

РОЗДІЛ 2

АНАЛІЗ ІНСТРУМЕНТІВ ДЛЯ ВАЛІДАЦІЇ ДАНИХ

2.1 Огляд популярних інструментів для валідації даних на ринку

У сучасному світі високоякісні дані є основою для прийняття обґрунтованих бізнес-рішень, ефективного управління та успішної роботи. Досягнення та підтримка високої якості даних вимагає використання спеціалізованих інструментів та методологій, які допомагають організаціям забезпечити точність, повноту, узгодженість, достовірність і актуальність даних, а також їх відповідність вимогам конкретного програмного забезпечення.

Далі наведено огляд найбільш популярних інструментів для валідації даних.

Перше місце з цього переліку посідає Astera.

Astera - це уніфіковане рішення для управління даними корпоративного рівня з розширеними функціями перевірки даних. Інструмент пропонує гнучкі можливості очищення та виправлення даних, надає доступ до перевірених, точних та узгоджених даних для отримання надійних аналітичних висновків.

Платформа також дозволяє впроваджувати суворі перевірки даних і налаштовувати правила відповідно до ваших конкретних вимог. Крім того, завдяки перевірці стану даних у режимі реального часу забезпечує миттєвий зворотній зв'язок щодо якості даних, що дозволяє відстежувати зміни.

Щодо інших особливостей, то візуальний та інтерактивний інтерфейс Astera дозволяє просто перетягувати об'єкти, маніпулювати даними та записувати їх у вибране місце без написання жодного рядка коду. Також Astera дозволяє легко підключатися до локальних систем, а також до хмарних джерел. Більше того, можна створювати підключення на основі API для будь-якої інтеграції або імпортувати їх з бібліотеки користувацьких підключень Astera.

Загалом це потужний інструмент для управління даними корпоративного рівня, що пропонує розширені можливості перевірки та очищення даних, забезпечуючи надійність, швидкість та зручність валідації даних.[6]

Наступний інструмент, з яким варто ознайомитись, це Informatica Data Quality. Це платформа для управління даними, яка дозволяє користувачам виконувати критичні завдання щодо якості даних, такі як дедуплікація, стандартизація, збагачення та валідація. Користувачі можуть виявляти, виправляти та відстежувати проблеми з якістю даних у хмарному середовищі.

Щодо функціональних особливостей, то платформа дозволяє користувачам підключатися до різних вихідних систем, таких як бази даних, файлові системи або додатки на основі SaaS.

Інструмент дозволяє користувачам профілювати, стандартизувати та перевіряти дані за допомогою попередньо створених правил та виконуваних оточень, а рішення надає користувачам видимість набору даних для виявлення та ідентифікації будь-яких розбіжностей.

Informatica дозволяє користувачам виконувати кілька завдань одночасно, розбиваючи їх на частини, забезпечуючи менші витрати часу.

Отже, Informatica Data Quality це не менш потужний інструмент, який дозволяє швидко та оперативно працювати з даними, надаючи достатній перелік переваг користувачам.[6]

Наступний інструмент, який ми розглянемо стане Talend.

Talend - це платформа для забезпечення якості та інтеграції даних, що надає користувачам доступ до точних даних. Talend профілює, очищує та стандартизує дані в різних системах. Завдяки алгоритмам машинного навчання інструмент пропонує рекомендації щодо виправлення та усунення проблем з якістю даних. Крім того, вбудований Trust Score забезпечує оцінку загального стану даних для виявлення розбіжностей і порушень у наборі даних.

Також є інтеграція даних для самообслуговування. Платформа самообслуговування для інтеграції даних Talend дозволяє користувачам створювати і розгортати завдання з інтеграції даних без написання коду.

Talend пропонує широкий спектр можливостей трансформації даних, включаючи фільтрацію, сортування, агрегування та об'єднання.

Також дозволяє користувачам готувати дані, застосовувати перевірки якості та відстежувати стан даних за допомогою Talend Trust Score. Дозволяє користувачам захищати конфіденційну інформацію, надаючи доступ на основі ролей, і забезпечує відповідність нормативним актам.[6]

Наступний інструмент це Datameer. Datameer - це рішення для підготовки та перетворення даних, яке перетворює необроблені дані у формат, придатний для аналізу. Платформа розроблена для Snowflake, хмарного рішення, що керує всіма аспектами життєвого циклу даних, від розвідки до підготовки та обміну перевіреними наборами даних, інженери та аналітики можуть трансформувати дані безпосередньо за допомогою простого SQL-коду або інтерфейсу без коду.

Пропонуючи інтерфейс у вигляді електронної таблиці, платформа дозволяє користувачам добре орієнтуватися та взаємодіяти зі складними даними.

Крім того, функція кодування Datameer автоматично перетворює категоріальні дані в двійковий формат для використання в моделях машинного навчання, а функція каталогу забезпечує централізоване представлення всіх активів даних в організації з можливостями інтелектуального пошуку.[6]

Ще один інструмент, який допоможе перевірити дані на точність, повноту та відповідність визначеним стандартам це Alteryx. Це платформа для підготовки даних та аналітики, яка забезпечує доступ до своєчасної інформації. Вона працює як конструктор робочих процесів, пропонує підключення до різних джерел, таких як плоскі файли, підключення до баз даних, API тощо. Платформа використовує штучний інтелект для надання рекомендацій щодо

покращення якості даних, що допомагає перевіряти, трансформувати та фільтрувати дані відповідно до вимог.

Alteryx пропонує можливості профілювання даних, які дозволяють користувачам зрозуміти характеристики даних і виявити потенційні проблеми. Платформа дозволяє користувачам відстежувати та керувати даними, вести журнали аудиту та забезпечувати контроль доступу, пропонує функції моніторингу продуктивності, які можна використовувати для відстеження ефективності робочих процесів підготовки даних та аналітики.[6]

Data Ladder - це рішення для контролю якості даних із вбудованими функціями профілювання, очищення та дедуплікації даних. Програмне забезпечення перевіряє дані перед тим, як зберігати їх у базі даних, пропонуючи перевірку якості даних у режимі реального часу. Платформа пропонує візуальний та інтерактивний інтерфейс, що дозволяє технічним бізнес-користувачам обробляти дані в середовищі без коду.[6]

Інструмент дозволяє користувачам знаходити, очищати, перевіряти і зіставляти дані відповідно до конкретних вимог бізнесу до якості даних, а також інтегрувати дані з різних джерел, включаючи формати файлів, реляційні бази даних, хмарні сховища та API.

OpenRefine, раніше відомий як Google Refine, є потужним інструментом з відкритим вихідним кодом, який надає можливості для роботи з великими та складними наборами даних. OpenRefine дозволяє обробляти пропущені або некоректні значення, підтримує перетворення даних з одного формату в інший. Користувачі можуть виконувати масові зміни у масивах даних, такі як розбиття та об'єднання стовпців, зміна формату тексту та інше, що дозволяє легко адаптувати дані до вимог конкретних завдань.

Використання API забезпечує можливість підключення до сторонніх сервісів та автоматизації операцій, для збагачення та узгодження інтеграція дозволяє пов'язувати дані з зовнішніми базами даних. Інструмент підтримує експорт у різні формати, включаючи CSV, JSON, XML, EXCEL та інші.

OpenRefine працює на операційних системах Windows, Mac та Linux, що забезпечує широку підтримку платформ. Активна спільнота користувачів та розробників забезпечує підтримку та розвиток інструменту, роблячи його ще більш потужним та корисним.[7]

Загалом розгляд та аналіз різних інструментів для обробки даних показав, що на ринку існує широкий вибір рішень, які задовольняють потреби користувачів у різних аспектах управління даними.

Кожен з розглянутих інструментів має свої унікальні переваги та особливості, що дозволяють користувачам вибирати найбільш підходяще рішення для своїх конкретних потреб. Використання цих інструментів допомагає підвищити якість даних, покращити процеси обробки та аналізу, а також забезпечити точність та узгодженість даних у різних системах та платформах.

2.2 Методика тестування

Для того, щоб якісно і об'єктивно оцінити інструменти для валідації даних для програмного забезпечення у сфері телекомунікацій необхідна певна і чітка методика роботи. Іншими словами систематичний підхід до оцінки інструментів з метою визначення їхньої працездатності, ефективності та відповідності вимогам. В рамках цієї методики проводитиметься ретельне тестування інструментів на тестових даних, що відображають сценарії та варіанти використання в телекомунікаційній галузі. Результати тестування допоможуть визначити, які інструменти найбільш ефективно відповідають вимогам галузі та можуть бути рекомендовані для використання в практичних завданнях.

Перший необхідний крок, який допоможе оцінити і порівняти інструменти це створення тестових даних: згенеровані тестові дані для цієї методики допоможуть відобразити різні сценарії та варіанти використання в

галузі телекомунікацій. Наприклад, датасет “Інформація про абонентів” включатиме персональні дані, такі як ім'я, адреса, номер телефону, а також технічні дані про підключення або послуги, що використовуються. Враховуючи те, що згенеровані дані близькі до реальних можна протестувати інструменти для валідації даних на реальному прикладі. Тобто перевірити, чи виявляють інструменти пропущені або некоректно внесені дані, чи зможуть вказати та розрізнити ці помилки.

Наступним кроком даної методики є визначення критеріїв для оцінки інструментів. Тобто необхідно сформулювати конкретні критерії, за якими буде оцінюватись кожен інструмент. Ці критерії можуть включати швидкодію, надійність, масштабованість та інші аспекти, які є важливими для сфери телекомунікацій. Для зручного відображення оцінки кожного інструменту я обрала табличний формат, що здається мені найоптимальнішим для даної методики.

Третій крок, це поверхнева оцінка функціоналу вибраних інструментів. Тобто швидкий перегляд і загальне оцінювання основних можливостей та характеристик інструменту без детального вивчення. Цей спосіб дозволяє швидко отримати уявлення про те, чи відповідає інструмент загальним потребам і чи варто розглядати його для подальшого використання. Поверхнева оцінка може включати такі критерії як: оцінка доступних функцій для валідації даних, огляд інтерфейсу користувача для зручної навігації і використання інструменту, перевірка підтримуваних джерел і ефективність їх інтеграції, здатність інструменту працювати з великими обсягами даних та продуктивність обробки, перегляд та підтримка валідаційних правил.

Така поверхнева, але ефективна оцінка дозволить швидко визначити, наскільки інструменти відповідають загальним потребам у валідації даних і чи потрібно проводити більш детальний аналіз їхніх можливостей.

Наступний етап є одним з ключових, це обробка тестових даних вибраними інструментами. Цей процес включає використання різних функцій

інструментів для профілювання, очищення, стандартизації, дедуплікації та валідації даних. Мета профілювання даних полягає в ідентифікації потенційних проблем у масиві даних перед початком їх обробки. Очищення включає в себе виправлення помилок у даних, заповнення відсутніх значень та інші дії для покращення якості даних. Стандартизація - перетворення даних у єдиний формат, що відповідає вимогам телекомунікаційної галузі. Дедублікація – процес виявлення та видалення дублікатів з даних для забезпечення їх унікальності. Валідація - перевірка відповідності даних певним правилам і вимогам.

Після обробки тестових датасетів за допомогою кожного інструменту заповнюється таблиця з визначеними критеріями оцінки та проводиться аналіз результатів.

За допомогою описаної методики я планую порівняти вибрані інструменти між собою і виділити переваги та недоліки кожного з них, також за допомогою описаних кроків можна систематично тестувати та порівнювати різні інструменти для валідації даних у контексті сфери телекомунікацій, забезпечуючи об'єктивність та достовірність отриманих результатів.

2.3 Критерії оцінки інструментів валідації даних

Критерії оцінки інструментів валідації даних визначають ефективність та надійність процесу валідації, включаючи різні аспекти, які впливають на загальну якість та точність даних, залежно від типу даних, з якими працює користувач та призначенням інструменту.

Для забезпечення об'єктивного порівняння та аналізу переваг та недоліків різних інструментів валідації даних я обрала такі ключові аспекти:

- Personal Impression:

a. Usability(зручність використання) - визначає наскільки легко і ефективно користувачі можуть взаємодіяти з інструментом для досягнення своїх цілей.

Інструмент повинен бути зрозумілим для користувачів з різним рівнем технічної підготовки. Інтуїтивно зрозумілий інтерфейс дозволяє швидко освоїти роботу з інструментом без необхідності тривалого навчання, а чітке і логічне розміщення функцій і опцій робить інструмент простішим у використанні і сприяє зниженню кількості помилок та підвищенню продуктивності користувачів. Важливо, щоб інструмент був розроблений з урахуванням зручності роботи, включаючи можливість персоналізації налаштувань, зручного відображення результатів та можливості швидкого доступу до потрібних функцій.

b. Response Time, час реакції інструменту на дії користувача, включаючи завантаження сторінок та виконання операцій, а також загальну продуктивність системи.

Інструмент повинен швидко завантажувати основні та додаткові сторінки своєї платформи, крім того, час, необхідний для виконання основних операцій, таких як імпорт, очищення, профілювання та валідація даних має бути оптимальним для зручної роботи користувача, адже висока швидкість обробки даних є критично важливою для підтримки високої продуктивності.

c. Help and Support, визначає доступність підтримки користувачів, можливість отримання допомоги у разі виникнення питань або проблем.

Важливо, щоб при користуванні інструментом завжди був доступ до технічної підтримки через різні канали зв'язку такі як електронна пошта, телефон, онлайн-чат або спеціалізовані форуми. Навчальні матеріали також є невід'ємною частиною підтримки користувачів: інструкції, відео-уроки або технічні посібники. Також оновлення та покращення, вагомою перевагою є те, коли розробники інструменту регулярно оновлюють та

покращують інструмент, враховуючи зворотній зв'язок від користувачів.[8]

- Accessibility:

a. Offline tool, можливість використання інструменту без підключення до Інтернету.

Інструмент, який може працювати офлайн, не втрачаючи своїх функціональних можливостей має вагомі переваги, включаючи гнучкість у використанні та підвищення безпеки даних.

b. Trial period, тривалість пробного періоду для оцінки інструменту перед придбанням.

Важливо, щоб кількість днів пробного періоду була достатньою аби випробувати всі основні функції та переконатися, що інструмент відповідає визначеним потребам і вимогам.

c. Price, вартість ліцензії, плата за обслуговування та оновлення, додаткові витрати.

Важливо, щоб вартість інструменту відповідала якості та користі від його використання.

d. Accessibility Feature, доступність для користувачів з обмеженими можливостями.

Доступність для користувачів з обмеженими можливостями є важливим критерієм для забезпечення інклюзивності та рівного доступу до інструменту. До цього списку можуть входити екранні читачі, можливість налаштування інтерфейсу для користувачів із вадами зору або слуху, та забезпечення легкого навігаційного досвіду.

e. Impressive Feature, особлива функція або можливість інструменту, що вигідно відрізняє його від інших.

Особливою функцією інструменту можуть бути унікальні методи обробки даних, інтеграція з іншими системами, покращена безпека, автоматизоване виправлення помилок та інше. [8]

- Test Optimizations:

a. Test suite preconditions(single login), умова, при якій користувач має єдиний вхід (логін) до системи або додатку.

Ця умова означає, що тестовий інструмент має здатність виконувати всі необхідні тести без повторно входу в систему для кожного тестового випадку. Перевагами такої функції є скорочення загального часу виконання тестів та менша кількість запитів до системи, що знижує навантаження на сервери.

b. Scalability, здатність інструменту ефективно працювати з великими обсягами даних та при наявності декількох тестових сценаріїв.

Це є критично важливим аспектом, особливо для компаній та підприємств, які регулярно працюють з великими наборами даних та складними системами. Інструмент повинен ефективно працювати з масштабними даними без зниження продуктивності та підтримувати паралельне виконання тестових сценаріїв.

c. DDT(Data-Driven Testing), можливість виконувати тестові сценарії з використанням різних наборів вхідних даних.

Перевагами DDT є зниження ручної роботи та підвищення ефективності за рахунок автоматизованих тестів, крім того, тестування з різними наборами даних забезпечує більш повну перевірку системи, що дозволяє виявити більшу кількість потенційних помилок.[8]

- Tests:

a. Variety of Test Cases, різноманіття типів тестових кейсів та сценаріїв для покриття різноманітних аспектів даних.

Важливо, щоб інструмент підтримував різноманітні типи тестів, включаючи функціональні, нефункціональні, регресійні, навантажувальні та інші. Крім того інструмент повинен дозволяти користувачам створювати та налаштовувати нові тестові сценарії відповідно до специфічних вимог в різноманітних галузях.

b. Data Complexity Handling, здатність ефективно валідувати та обробляти складні структури даних та їх взаємодію.

Враховуючи те, що телекомунікаційні дані можуть включати складні структури, такі як вкладені таблиці, різні типи взаємозв'язків та динамічні зміни структури даних, важливо, щоб інструмент ефективно обробляв ці структури та мав можливість перевіряти коректність та узгодженість взаємозв'язків між різними частинами даних.

c. Data Integrity Verification, перевірка цілісності та достовірності даних під час їх обробки та зберігання.

Важливо, щоб інструмент забезпечував перевірку цілісності даних на всіх етапах їх обробки та зберігання, крім того, відстежував всі зміни внесені до даних, історія змін важлива для подальшого аудиту.[8]

- Tests Execution and Reporting:

a. Test creation time, швидкість створення тесту.

Для зручного і ефективного використання інструменту, важливо, щоб час для створення нового тестового сценарію або кейсу був оптимальним. Перевагою буде можливість повторного використання раніше створених тестів або їх частин для економії часу та підвищення ефективності тестування.

b. Test execution time, швидкість виконання тестів.

Швидкість виконання тестів включає час завантаження даних, обробку запитів та генерацію результатів. Інструмент повинен бути здатним обробляти великий обсяг даних без значного впливу на загальну продуктивність системи., особливо це важливо для телекомунікаційних систем, які працюють з великими обсягами даних у реальному часі.

c. Test report, точність та повнота звітності про результати виконаних тестів.

Після виконаних тестів необхідно бачити точні та детальні звіти про результати, включаючи виявлені помилки, тому важливо, щоб інструмент надавав повну інформацію про тестування.[8]

- Capabilities:

a. DB integration, інтеграція інструменту з базами даних.

Високо оцінюється здатність інструменту підключатися до різних типів баз даних, що дозволяє виконувати тестування даних безпосередньо з джерел їх зберігання, також здатність інструменту підтримувати виконання складних SQL-запитів для ефективної валідації великих обсягів даних

b. API integration, інтеграція інструменту з API для тестування.

Для того, щоб забезпечити широкий спектр можливостей інтеграції, інструмент повинен підтримувати різні протоколи API, для прикладу REST, SOAP або GraphQL. Перевагою буде можливість автоматизованого тестування API для забезпечення швидкості та ефективності процесу валідації даних.

c. Customization and extensibility, гнучкість інструменту у налаштуванні та розширенні для задоволення конкретних бізнес-вимог.

Крім створення спеціальних правил валідації та сценаріїв тестування, інструмент повинен підтримувати розширення своїх функцій за допомогою плагінів або додаткових модулів для задоволення унікальних потреб користувачів.

d. SDK, наявність SDK для розширення функціональності інструменту.

Перевагою інструменту буде наявність програмного комплексу розробника(SDK) для створення власних розширень і додатків, що дозволяє адаптувати інструмент до специфічних вимог телекомунікаційної галузі.

e. CI/CD integration, можливість інтегрувати інструмент з процесами безперервної інтеграції/безперервного розгортання.

Оцінює здатність підтримки популярних CI/CD інструментів таких як Jenkins, GitLab CI, Travis CI, CircleCI, Azure DevOps та інших, дозволяє легко інтегрувати процеси тестування даних у вже існуючі робочі процеси.

f. Data cleansing and transformation, очищення та перетворення даних;

Цей критерій оцінює здатність інструменту проводити очищення та перетворення даних для забезпечення їх точності, узгодженості та відповідності вимогам бізнесу. Інструмент повинен виявляти такі помилки як дублікати, пропущені значення, некоректні формати, та автоматично або напівавтоматично виправляти їх. Також цей критерій оцінює здатність об'єднувати дані з різних джерел та агрегувати їх для спрощення аналізу та забезпечення повноти інформації. Інструмент повинен дозволяти змінювати формат даних (наприклад, зміна формату дати) для забезпечення відповідності даних конкретним вимогам.

g. Data visualization, візуалізація даних.

Цей критерій оцінює здатність інструменту представляти дані у вигляді візуальних елементів, таких як інтерактивні інформаційні панелі, діаграми та графіки, можливість налаштовувати вигляд та параметри діаграм: кольори, розміри та мітки. Візуалізація даних є важливою для аналізу, моніторингу та прийняття рішень у будь-якій сфері.[8]

Висновки

У наш час на ринку існує багато ефективних рішень для валідації даних. Кожен з інструментів заявляє про безліч своїх функцій і можливостей, але для того, щоб перевірити чи дійсно вони ефективні, потрібно їх протестувати на

основі даних, що наближені до реальних. Для цього власне і потрібна методика, яка покроково дозволить дізнатись про інструмент більше і порівняти його з своїми конкурентами. Методика тестування є важливим елементом, що забезпечує об'єктивність і точність оцінки інструментів. Важливими є і критерії, за якими оцінюються інструменти, адже вони визначають ефективність та надійність процесу валідації, враховуючи різні аспекти, що впливають на загальну якість та точність даних.

Систематичний підхід до оцінки інструментів валідації даних, підтриманий чіткою методикою, є ключовим фактором у визначенні найбільш ефективних рішень, які відповідають специфічним потребам телекомунікаційної галузі.

РОЗДІЛ 3

ТЕСТУВАННЯ РОБОТИ ВИБРАНИХ ІНСТРУМЕНТІВ

3.1 Створення датасетів

Тестування роботи вибраних інструментів для валідації даних є найважчим і водночас найцікавішим етапом даного дослідження. Хоча ознайомитись і оцінити функціонал інструментів можна й теоретично, це не буде мати ніякого сенсу без практичного застосування. Для практичного тестування утиліт необхідно мати певні набори даних із окремої області. Саме тому важливою частиною тестування є створення датасетів.

Для тестування інструментів я вирішила використовувати саме згенеровані дані, і на це є декілька причин:

- Згенеровані дані дозволяють створювати більш точні сценарії з різноманітними варіаціями, що відображають різні випадки використання.
- Під час тестування інструментів на основі згенерованих даних можна краще відслідковувати помилки, невалідні значення і саму роботу інструменту.
- Створення генерованих даних дає можливість обрати потрібний обсяг інформації, формат або ж вигляд. Наприклад, відобразити одні і ті ж дані можна у вигляді схем SQL або ж таблиць Excel.

Враховуючи ці всі переваги, потрібно було обрати інструмент для генерації даних. Я обрала - Moskaroo. Moskaroo - це безкоштовна платформа, яка дозволяє швидко та легко завантажувати великі обсяги випадково згенерованих тестових даних на основі власних специфікацій користувачів, які потім можна завантажити безпосередньо у тестове середовище за допомогою різних форматів. Серед інших переваг вибраного інструменту - це можливість використання на основі браузера, тобто Moskaroo можна використовувати як веб-програму на будь-якому пристрої з підключенням до Інтернету, що робить її доступною та зручною для користувачів[9].

Mockaroo підтримує широкий набір форматів даних, таких як CSV, JSON, SQL, Tab-Delimited, Cassandra CQL, Firebase, IndluxDB, Custom, Excel, XML та DBUnit XML. [10]

Для того, щоб максимально якісно протестувати вибрані інструменти я згенерувала набори даних.

Датасети мають однакові поля, включаючи id користувача, його ім'я, прізвище, дату народження, адресу, номер телефону, пошту, тип підключення, тип послуги, тарифний план та статус підключення (рис. 3.1).

Field Name	Type	Options
ID	Row Number	blank: 0 %
First_name	First Name	blank: 0 %
Last_name	Last Name	blank: 0 %
Date of birth	Datetime	01/01/1940 to 06/02/2018 format: m/d/yyyy blank: 0 %
Address	Street Address	blank: 0 %
Phone Number	Phone	format: +## ### #### blank: 0 %
Email	Email Address	blank: 0 %
Connection type	Custom List	50 Mbps, 100 Mbps, 1 Gbp random blank: 0 %
Services	Custom List	Internet, Phone, Cable TV random blank: 0 %
Tariff plan	Custom List	Gold Plan, Premium Package, Basic Plan random blank: 0 %
Account Status	Custom List	Active, Blocked, Disconnected random blank: 0 %

Рис. 3.1 Створення датасету в програмі Mockaroo

Вихідний датасет містить 1000 рядків, це також значення яке можна регулювати в залежності від потреб. Я вважаю це оптимальною кількістю для даного дослідження, що дозволить відстежувати і контролювати роботу інструментів.

Зручною функцією інструменту є те, що можна переглянути, як виглядатимуть вихідні дані перед початком генерації.

На рис. 3.2 представлено валідний набір даних у форматі CSV.

Preview

	TABLE				RAW			
ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Service
1	Clevey	Rickett	6/25/1956	24653 Cherokee Center	+98 567 970 7747	crickett0@bravesites.com	1 Gbp	Cable
2	Molly	Spridgen	1/16/1965	93 Parkside Trail	+33 871 497 9776	mspridgen1@addthis.com	1 Gbp	Phone
3	Inna	McReidy	9/23/1949	761 Sommers Way	+55 751 825 5840	imcreidy2@hao123.com	1 Gbp	Intern
4	Marv	Lampkin	5/29/2006	298 Clyde Gallagher Drive	+47 845 414 3779	mlampkin3@hao123.com	100 Mbps	Phone
5	Zea	Burnell	2/14/1970	7 Summerview Point	+86 384 481 9975	zburnell14@facebook.com	50 Mbps	Cable
6	Wynn	Bussetti	1/15/1971	4 Jenna Park	+62 884 279 5573	wbussetti5@yelp.com	100 Mbps	Cable
7	Gertrud	Grosvener	7/25/1977	128 Sachs Junction	+86 895 625 0674	ggrosvener6@springer.com	100 Mbps	Cable
8	Maxy	Naptine	1/9/1991	774 La Follette Road	+7 213 385 6153	mnaptine7@yandex.ru	50 Mbps	Intern
9	Say	Garie	11/13/2005	3 Maywood Road	+33 306 734 4111	sgarie8@mozilla.com	50 Mbps	Phone
10	Nani	Louden	8/12/1967	60247 Warbler Court	+54 833 148 4540	nlouden9@king.com	100 Mbps	Intern
11	Constancy	Floris	5/30/1940	28713 Pennsylvania Junction	+33 167 792 6625	cflorisa@1688.com	1 Gbp	Cable
12	Guglielmo	Basili	8/9/2001	1 Claremont Plaza	+81 149 996 4629	gbasilib@guardian.co.uk	1 Gbp	Phone
13	Consuela	Garrettson	1/14/1996	08656 Bubler Lane	+63 693 862 7556	garrettson@bloomberg.com	100 Mbps	Cable

showing first 100 rows

Rows: 1000 GENERATE DATA CLOSE

Рис. 3.2 Попередній вигляд згенерованих даних у форматі CSV

Для того, щоб краще зрозуміти роботу інструментів і мати можливість відслідкувати їх роботу я вручну створила ряд помилок, посилаючись на стандарт якості даних ПЗ - ISO 25012.

Перший навмисно порушений критерій - це точність, точніше його два підатрибути: синтаксична точність та семантична. У датасеті я зробила два помилкових значення: неправильний регістр та невідповідність змісту.

Друга помилка, це неповнота значень, датасет не містить одного зі значень поля Account Status та Services.

Далі, щоб протестувати інструменти на здатність виявляти помилки, була зроблена помилка, яка порушує критерій узгодженості, значення поля Account Status написано українською та російською мовами замість англійської. Хоч дані синтаксично правильні, вони не відповідають критерію узгодженості.

Достовірність - ще один критерій за яким можна перевірити чи дані якісні, так як наші дані мають бути неякісними, одна з дат поля Date of birth взята із майбутнього і порушує критерій достовірності. А значення пошти написано без собачки (@) - непомітна, але значна помилка, яку інструмент повинен виявити.

Щоб перевірити здатність інструменту виявляти неактуальні дані, одне зі значень поля Date of birth написано як 01/06/1900, тобто застаріле для актуальних даних(якщо ж ці дані не зберігаються для якогось архіву або історії). При тестуванні інструменту це також має враховуватись. А друга помилка, яка зроблена для цього критерію, це значення неактуального тарифного плану, якого б вже не мало бути серед списку.

Критерій відповідності, так як це міра відповідності стандартам, то 6-значний номер замість 12-значного потенційно має стати проблемою, яку інструмент повинен виявити, така ж ситуація і з номером, що складається з 14 цифр.

Ефективність - критерій, який оцінює дані в тому вигляді, в якому їх обробка займає меншу кількість часу. Якщо замість цифр написані слова, це також дані, але вони неякісні.

Згенерувавши всі необхідні формати даних і зробивши один з сетів невалідним можна переходити до наступного етапу дослідження.

3.2 Перевірка якості даних вибраними інструментами

Після створення датасетів можна переходити до практичної частини дослідження. Для перевірки якості даних я обрала перелік таких інструментів: Astera, OpenRefine, Data Ladder, Alteryx, Datameer, Talend, Informatica Data Quality.

Першим на черзі йде Astera. Даний інструмент з опису заявляв про свій широкий функціонал та ідеальність при роботі з валідацією даних. Але проблеми з ним почались ще на початку отримання пробного доступу. Для реєстрації та отримання пробного доступу потрібно зв'язатись із службою підтримки та домовитись про зустріч. Після написання листа в службу підтримки я отримала одну відповідь, але на етапі домовлення про зустріч

служба підтримки люб'язно мене проігнорувала. Саме по цій причині я не змогла оцінити, наскільки якісним є даний інструмент.

Наступним є інструмент OpenRefine, який має відкритий вихідний код. Даний інструмент завоював велику популярність серед аналітиків даних із нестандартними запитами, може виконувати безліч різноманітних дій та має можливість сильно розширювати функціонал, завдяки стороннім розширенням. Остання особливість і є тією причиною, чому даний інструмент залишився без моєї оцінки - початковий функціонал інструменту без сторонніх розширень має менше можливостей, аніж Microsoft Office Excel, який майже кожен з нас використовував, коли мав справу з таблицями, а оцінювати функціонал інструменту зі сторонніми розширеннями не має сенсу, так як функціонал, який вносять дані розширеннями, не являється стандартним і може не підтримуватись в новій версії інструменту.

З інструментом Datameer склалась ситуація схожа на ситуацію з інструментом Astera - доступ до пробного періоду видається лише при особистій зустрічі з менеджером по продажам, але після того як була залишена заявка на зворотній зв'язок, про зустріч так і не вийшло домовитись, і від даного інструменту на електронну пошту надходив лише спам.

Із інструментом Talend ситуація склалась цікавіше. Реєстрація на пробний період там відбувається автоматичне, і можна тестувати інструмент впродовж 14 пробних днів. Та все склалось не так ідеально, як мало б бути: для користувачів із пробним періодом доступ до інструменту надається тільки в ті періоди часу, коли навантаження на даний інструмент невелике. Впродовж 14 пробних днів я намагалась спробувати протестувати даний інструмент, але всі спроби були марними, так як навантаження на інструмент завжди було великим.

Далі на оцінці Data Ladder. Даний інструмент надає можливість підписатись на пробний період тривалістю в 30 днів. Для реєстрації потрібна електронна пошта з корпоративного домену, в моєму випадку підійшла електронна пошта з домену Ill.kpi.ua. Після реєстрації я отримала електронний

лист із персональним посиланням для завантаження інструменту. Даний інструмент підтримує лише платформу Windows.

Інтерфейс програми складається із двох основних частин: зліва позначені основні етапи обробки даних, а справа розміщена робоча область на якій буде відбуватись вся взаємодія.

Першим етапом є створення проекту (рис. 3.3). При створенні профілю можна вказати його ім'я та опис.

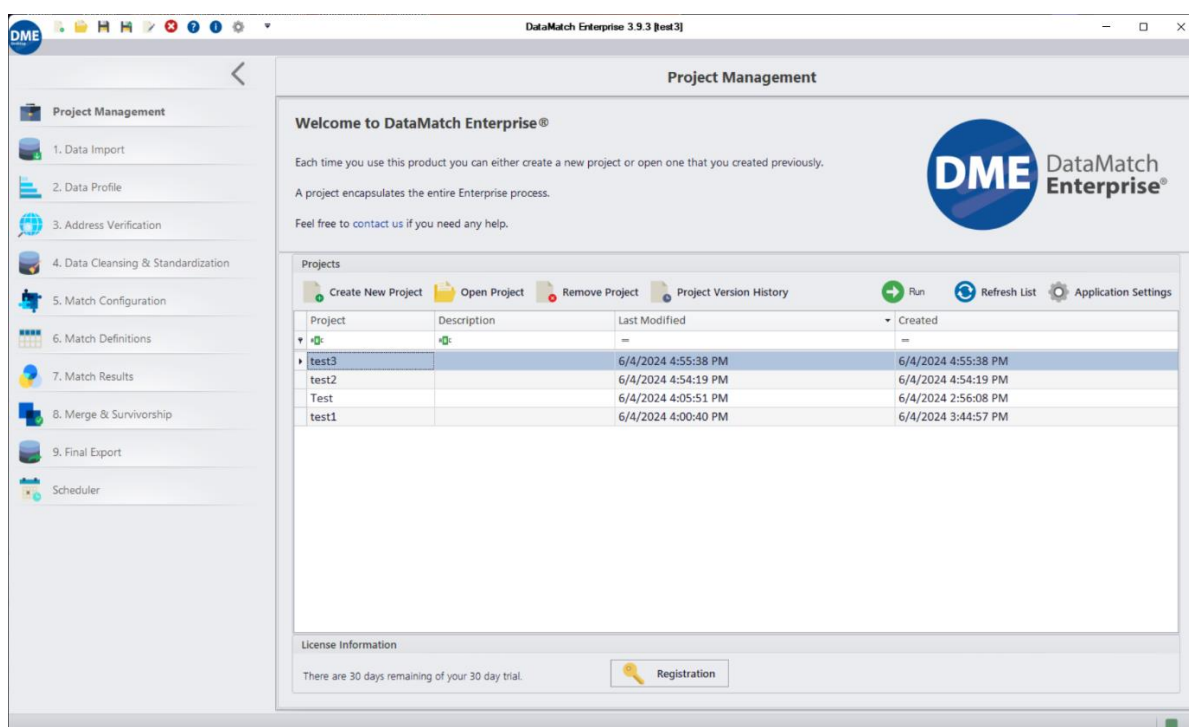


Рис. 3.3 Вікно створення профілю

Після створення проекту йде етап імпорту даних. На даному етапі можна обрати нове джерело даних, таке як файл, підключення з БД, підключення з хмарним сховищем, підключення з CRM та підключення з БД через драйвер ODBC, якщо потрібної БД немає в переліку. Перелік можливих джерел зображено на рисунках 3.4-3.8.

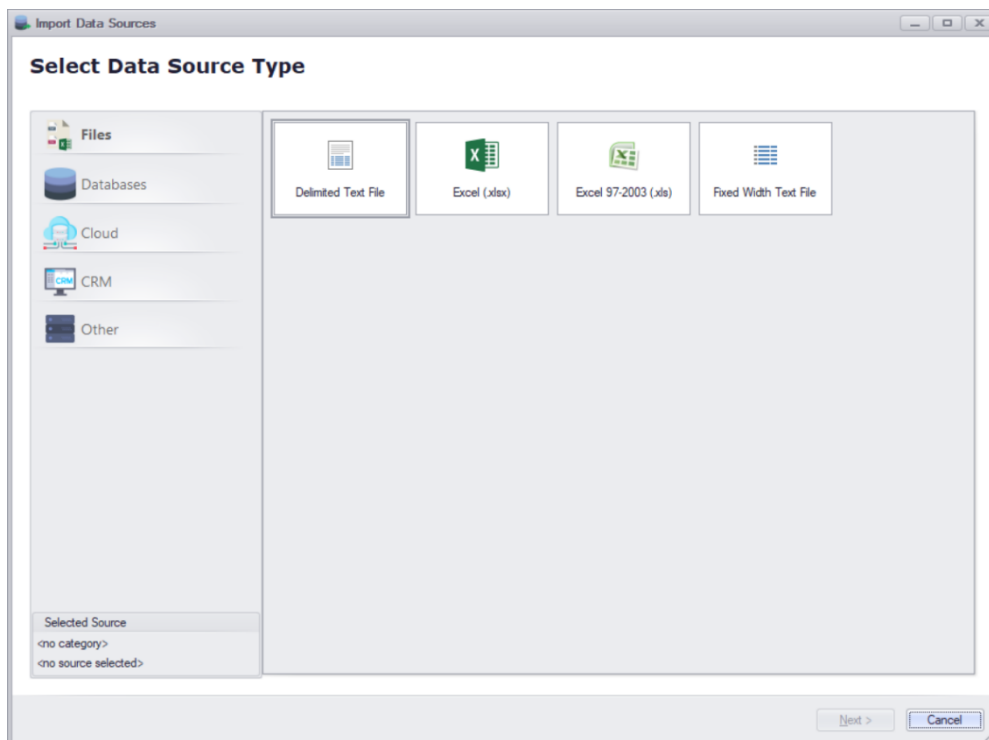


Рис. 3.4 Можливі варіанти завантаження з файлів

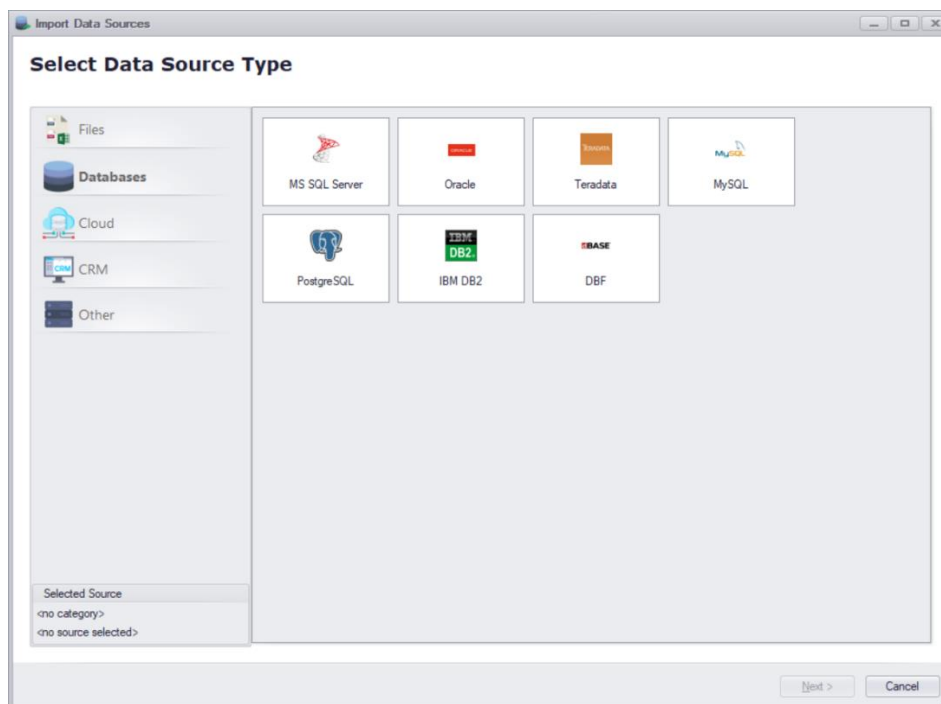


Рис. 3.5 Можливі варіанти завантаження з БД

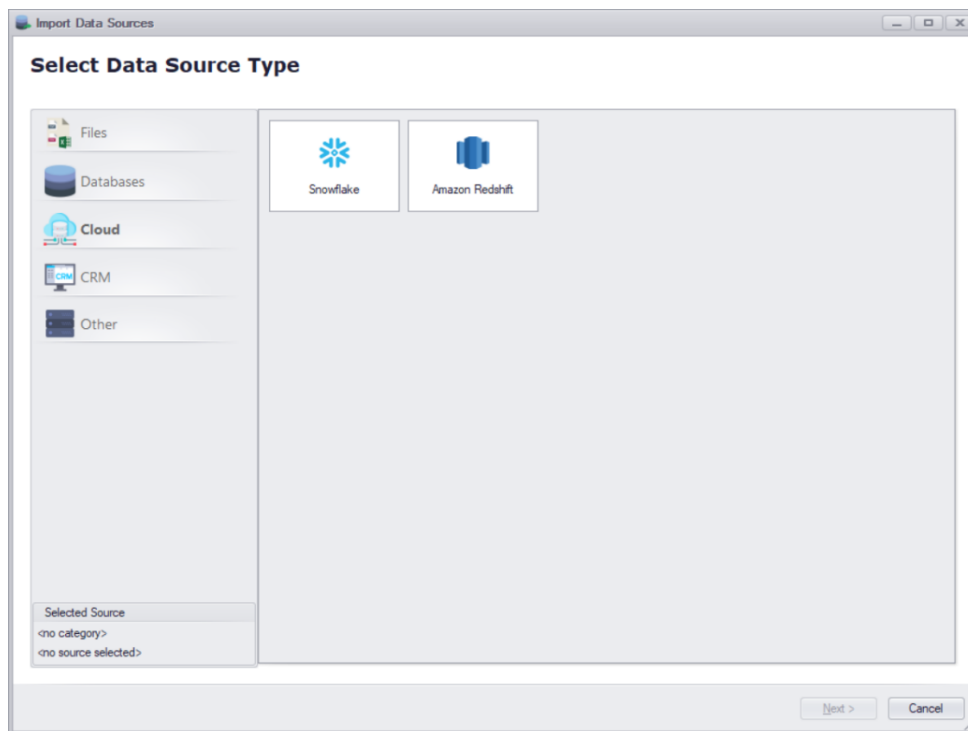


Рис. 3.6 Можливі варіанти завантаження з хмар

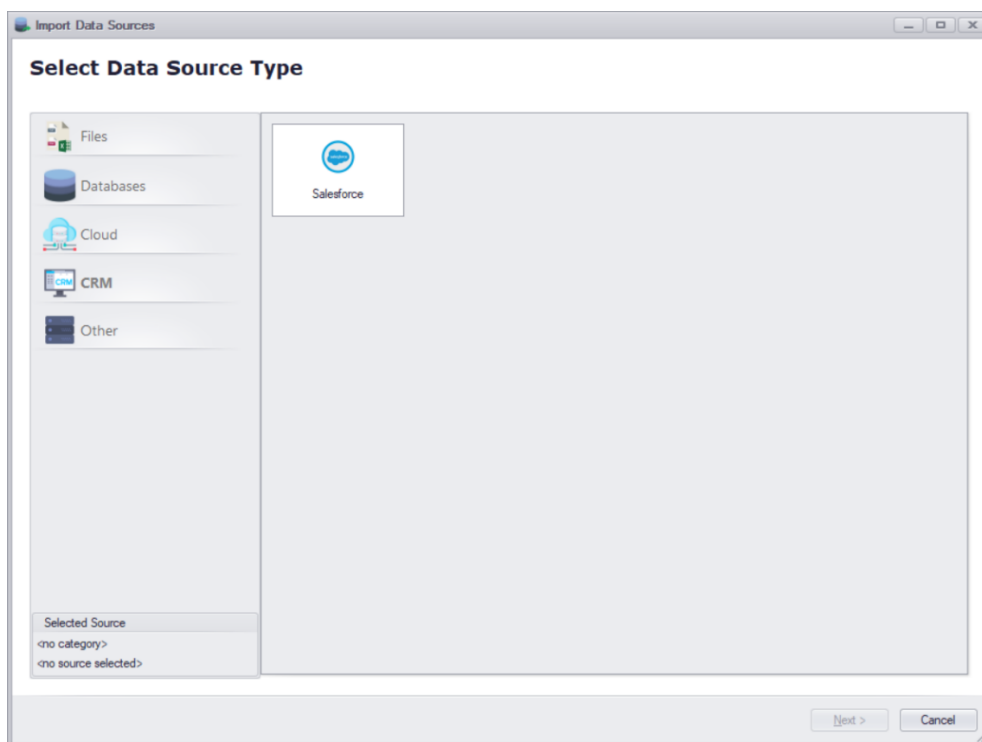


Рис.3.7 Можливі варіанти завантаження з CRM

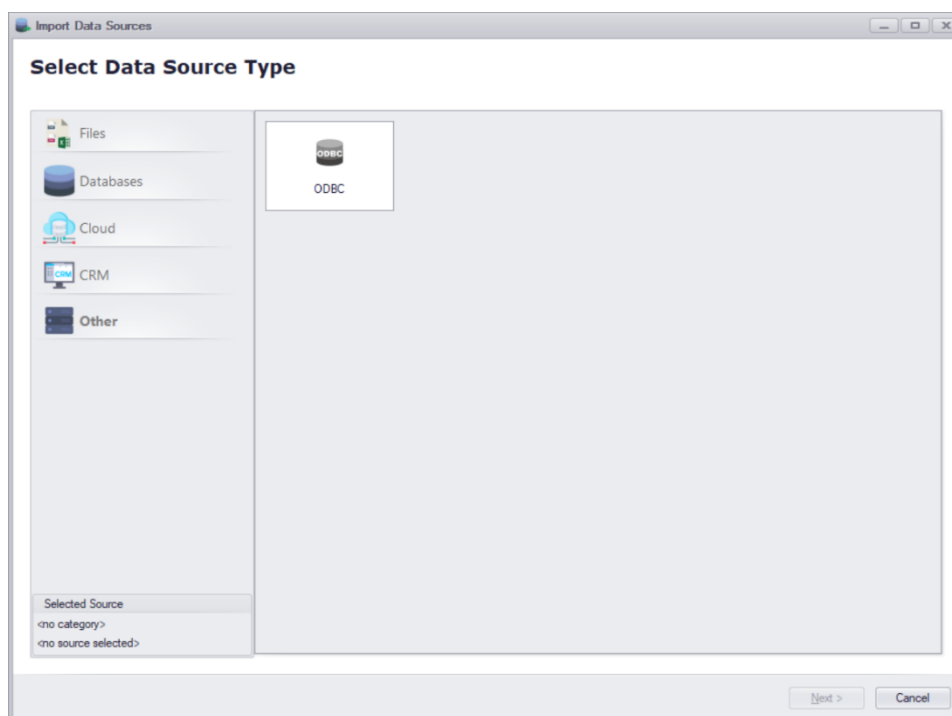


Рис.3.8 Можливі варіанти завантаження з інших джерел

Також є можливість обрати в якості джерела даних раніше використані джерела (рис.3.9).

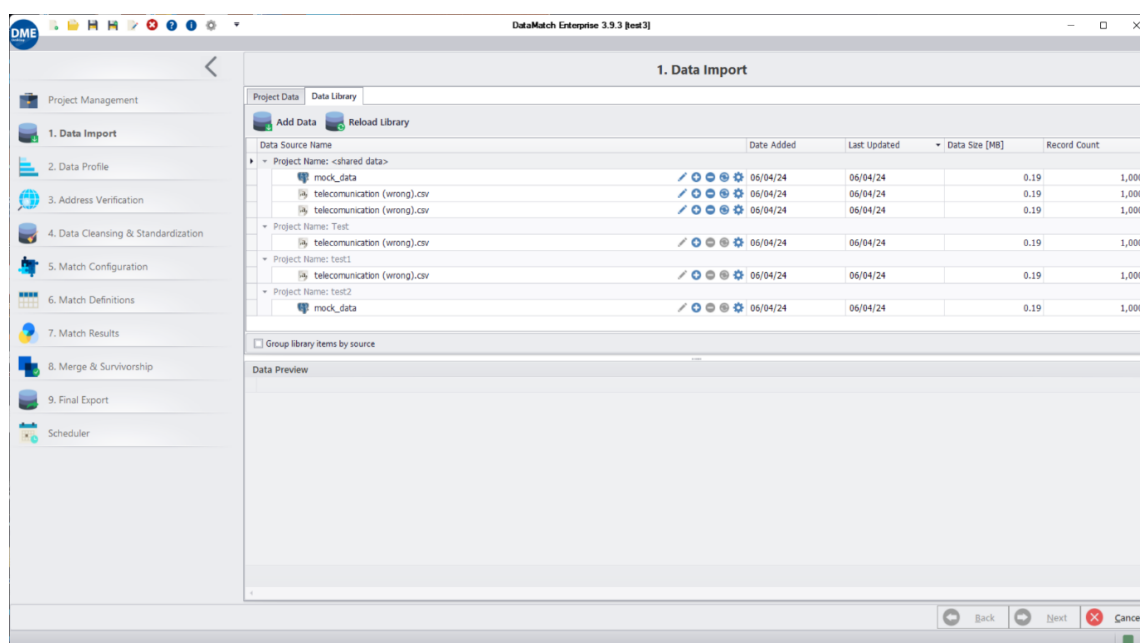


Рис.3.9 Вікно програми Data Import

Наступним кроком обробки даних є генерація профілю даних (рис.3.10). Під час цього етапу програмне забезпечення аналізує дані на певні патерни в кожному стовпці, на заповненість даного стовпця даними, максимальну довжину даних в цьому стовпці, кількість пустих полів, кількість унікальних значень, розбиття по символах, а також різні статистичні дані, такі як мінімальні та максимальні значення, мода, медіана і так далі. Ці дані є корисними для того щоб поверхнево оцінити якість даних, тобто якщо буде велика кількість пустих полів або полів типу Null, тоді це не якісні дані. Також за допомогою статистичних даних, такі як мінімальні та максимальні значення, можна оцінити, чи не виходять дані з дозволеного числового діапазону.

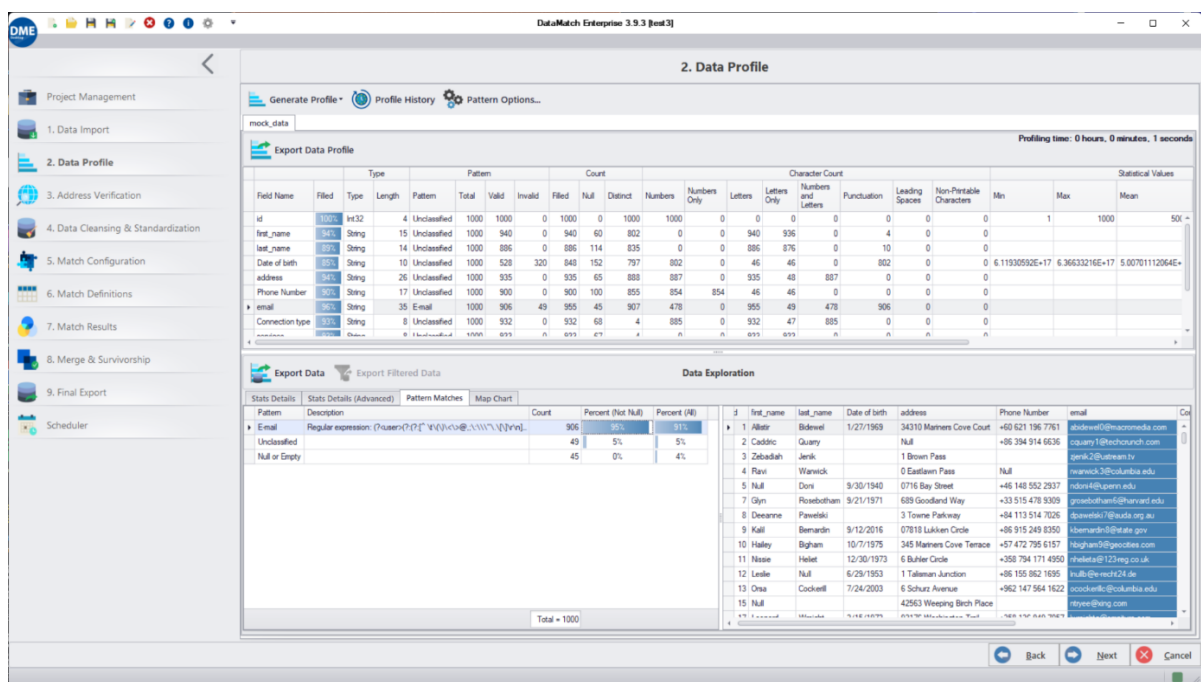


Рис.3.10 Вікно програми Data Profile

Наступним етапом з обробки даних в даному програмному забезпеченні є верифікація адрес, але вона доступна тільки для тих, в кого оформлена платна підписка (рис. 3.11).

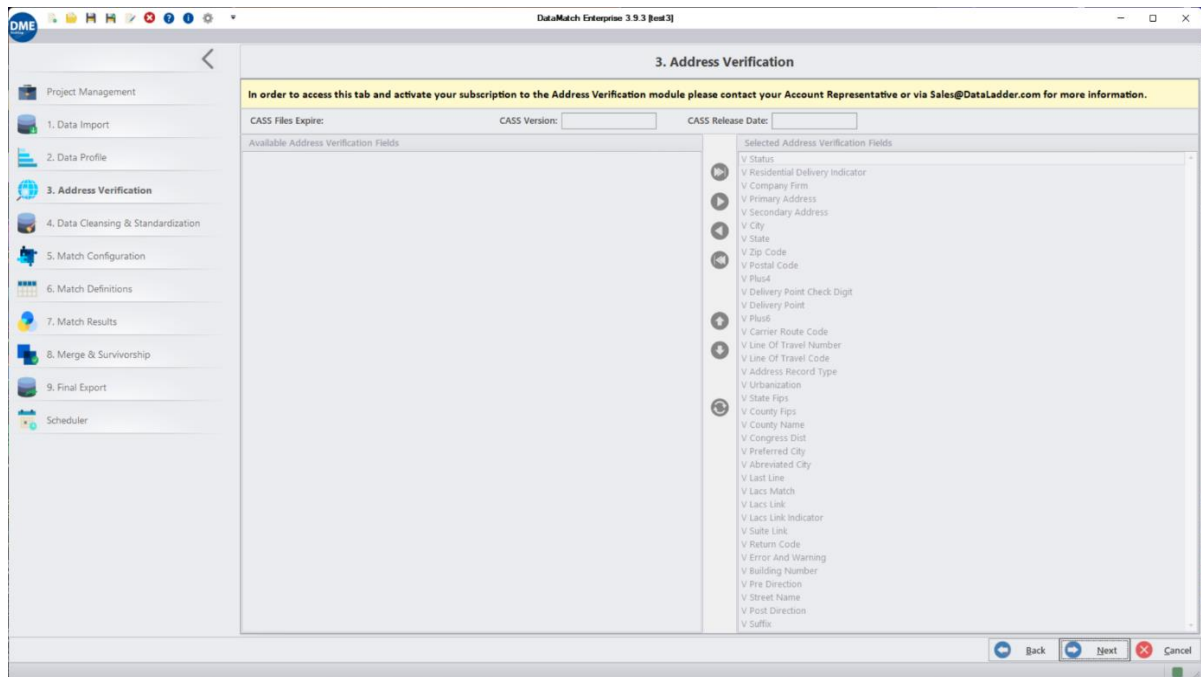


Рис.3.11 Вікно програми Address Verification

Наступним етапом є очищення та стандартизація даних (рис.3.12). На даному кроці інструмент надає можливість вказати, який регістр літер повинен використовуватись для певного поля, можна видалити або замінити символи, які неможливо ввести з клавіатури, також є можливість вказати значення для кожного поля, також можна видаляти або замінювати окремі символи, видаляти зайві пробіли на початку та в кінці значень, видаляти з певних значень пробіли, цифри або літери, якщо очікується, що дане поле не повинно містити цифр, пробілів або літер, також можна замінити цифру 0 на букву O, або навпаки - це може виявитись корисним, якщо ведеться робота з текстом, який розпізнавався на фотографіях або сканах документів, так як часто системи оптичного розпізнавання символів (англ. optical character recognition, OCR) путають дані символи. Вигляд даних інструментів зображено на рисунку

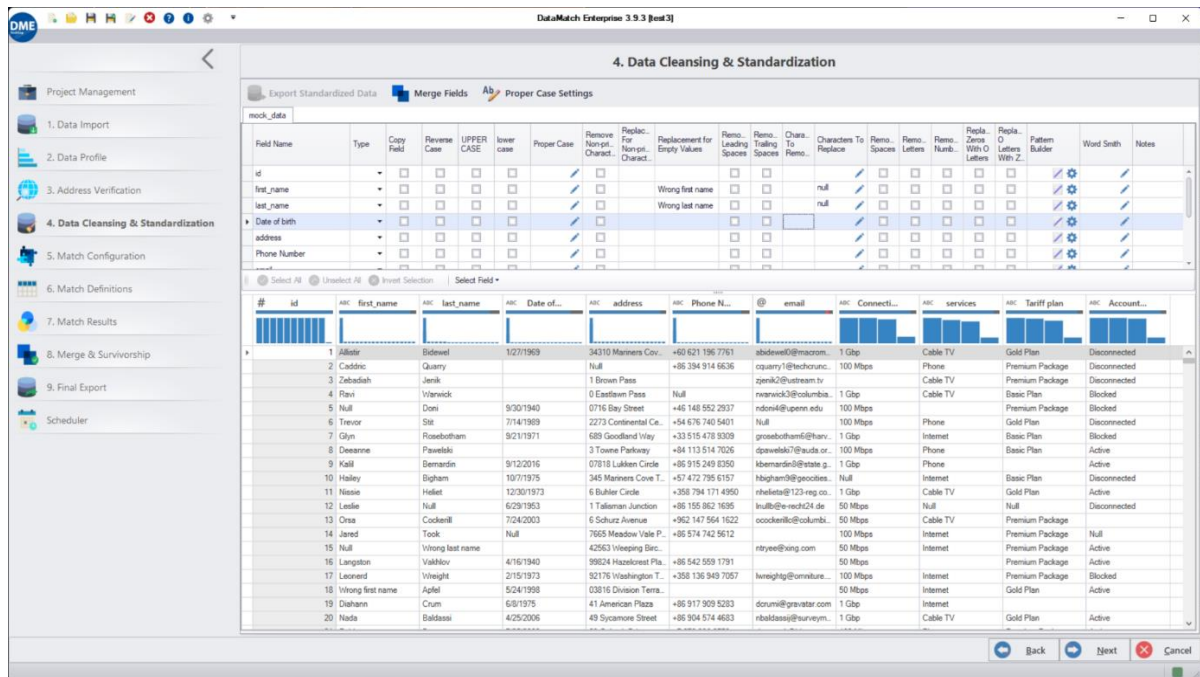


Рис.3.12 Вікно програми Data Cleansing & Standardization

Наступним кроком, який пропонує даний інструмент являється співставлення даних в різних джерелах. Даний інструмент може бути корисним у випадках, коли у вас є декілька файлів, які ви хочете об'єднати, і задля уникнення дублікатів при прямому копіюванні, що також потрібно для гарантування якості даних, інструмент знаходить однакові поля й відображає їх, після чого пропонує дії, наприклад видалення з першого джерела і в цьому випадку дані залишаться в другому файлі, або навпаки видалення з другого файлу і залишання даних в першому. Вигляд даного інструменту показаний на рисунку (рис.3.13).

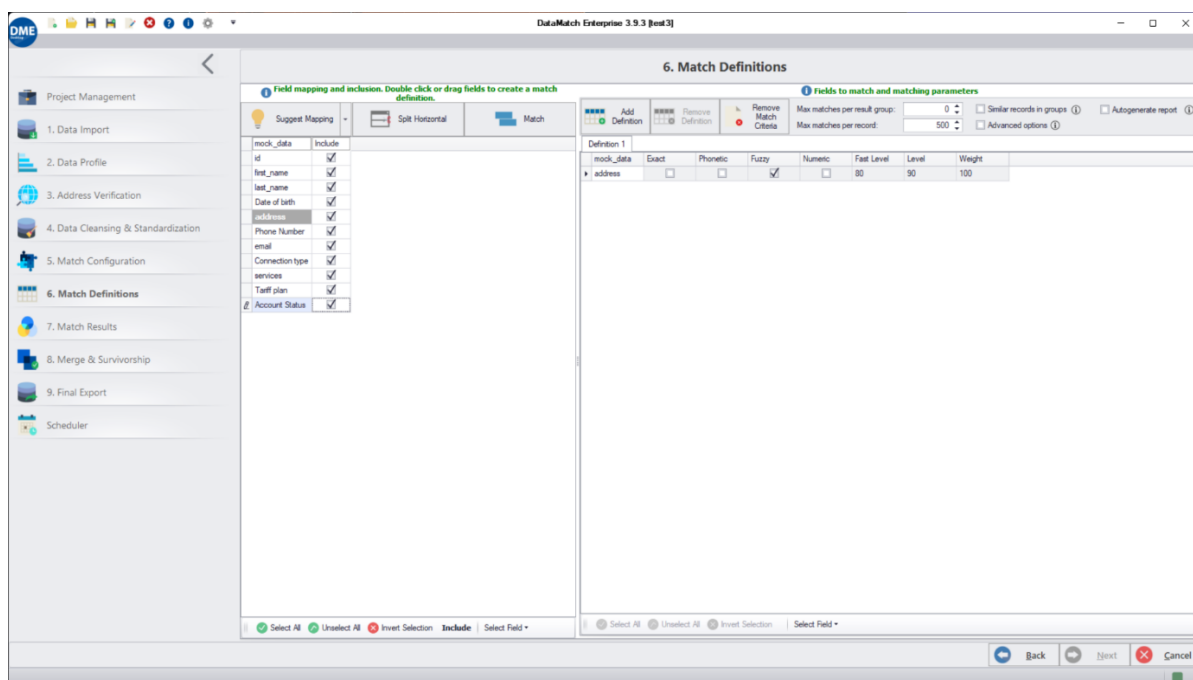


Рис.3.13 Вікно програми Match Definitions

Після цього кроку можна побачити результати збігів та зберегти їх в зручному для читання звіті. Після цього дані з різних джерел що співпали можна злити разом, при цьому вказавши, які дані і з якого джерела брати (рис.3.14).

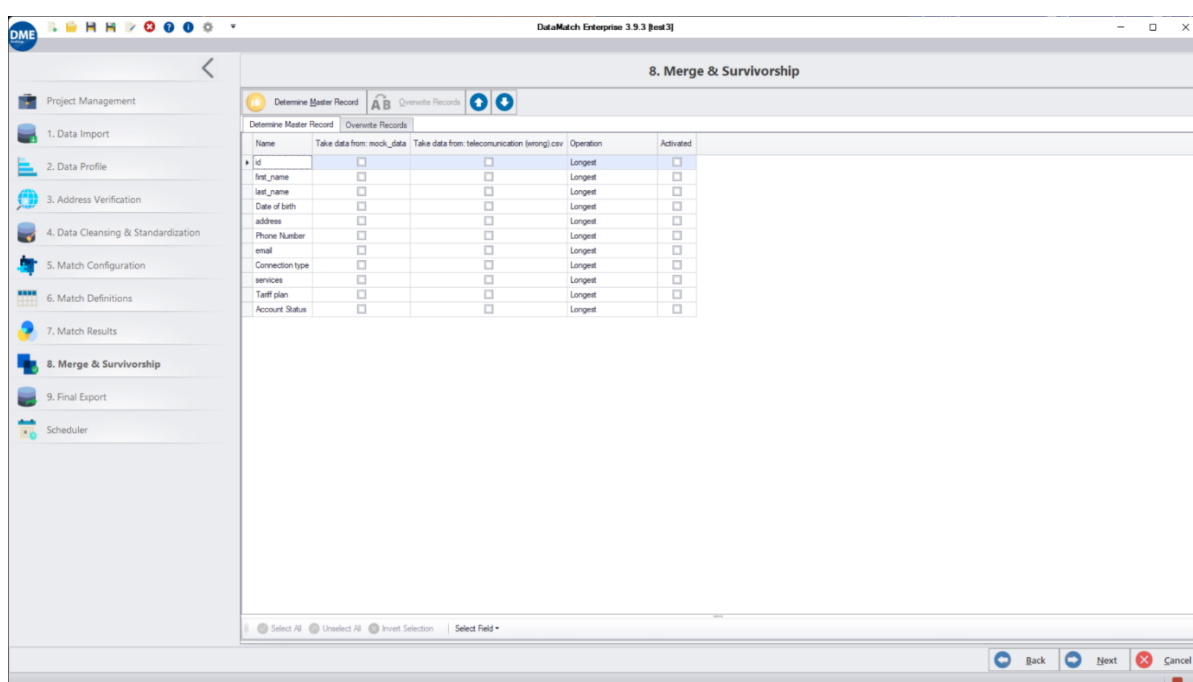


Рис.3.14 Вікно програми Merge & Survivorship

Фінальним кроком являється експорт отриманих результатів (рис. 3.15). В цьому модулі можна обрати, які дані потрібно експортувати, як поводитись із дублікатами полів, також можна зберегти лише дублікати. На жаль, в пробній версії даної програми відсутня можливість експортувати дані й перевірити, в які формати можна виконати експорт, а перевірити результат роботи програми можна лише на прев'ю.

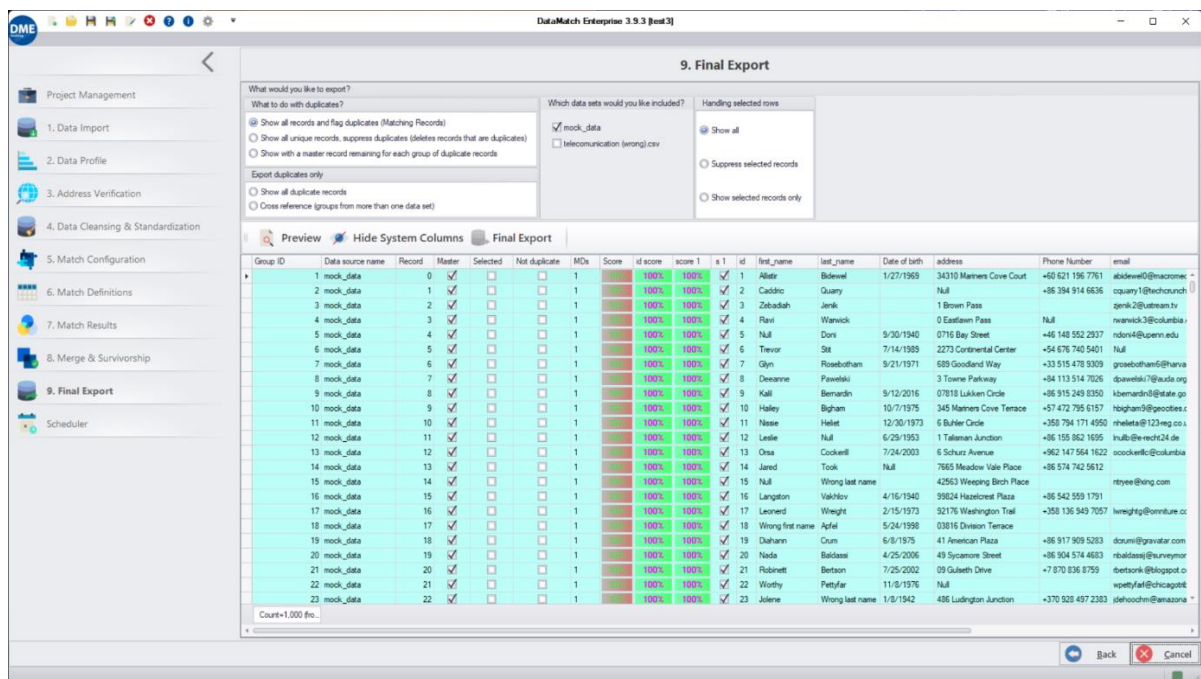


Рис.3.15 Вікно програми Final Export

Також в даному інструменті є модуль “Планувальник” в якому можна запланувати, щоб попередньо налаштовані виконувались із певною періодичністю, але знову ж таки, даний модуль не доступний в пробному періоді.

Перейдімо до тестування можливості налаштувати даний інструмент так, щоб він виявив помилки в датасеті з помилками, та по можливості виправив їх. Першим кроком імпортуємо дані з нашого датасету в даний інструмент, після

його проводим профілювання даних. На цьому етапі ми можемо оцінити загальну статистику по даним із датасету (рис.3.16)

Field Name	Filled	Type		Pattern			Count			Character Count								
		Type	Length	Pattern	Total	Valid	Invalid	Filled	Null	Distinct	Numbers	Numbers Only	Letters	Letters Only	Numbers and Letters	Punctuation	Leading Spaces	Non-Printable Characters
ID	100%	String	4	Unclassified	1000	1000	0	1000	0	1000	1000	1000	0	0	0	0	0	0
First_name	100%	String	12	First name	1000	745	255	1000	0	947	0	0	1000	999	0	1	0	0
Last_name	100%	String	13	Unclassified	1000	619	381	1000	0	990	0	0	1000	992	0	8	0	0
Date of birth	100%	String	10	Unclassified	1000	610	390	1000	0	986	1000	0	0	0	0	1000	0	0
Address	100%	String	29	Unclassified	1000	1000	0	1000	0	1000	1000	0	1000	0	1000	0	0	0
Phone Number	100%	String	21	Unclassified	1000	1000	0	1000	0	1000	1000	1000	0	0	0	0	0	0
Email	100%	String	36	E-mail	1000	999	1	1000	0	1000	529	0	1000	0	529	1000	0	0
Connection type	100%	String	16	Unclassified	1000	1000	0	1000	0	5	998	0	1000	2	998	0	0	0
Services	100%	String	10	Unclassified	1000	999	0	999	1	4	0	0	999	999	0	0	0	0
Tariff plan	100%	String	15	Unclassified	1000	1000	0	1000	0	4	0	0	1000	1000	0	0	0	0
Account Status	100%	String	12	Unclassified	1000	711	288	999	1	5	0	0	999	999	0	0	0	0

Рис.3.16 Загальна статистика даних із датасету

Даний інструмент має набір стандартних шаблонів, за якими може визначати тип даних в полі, і після цього аналізувати, чи є в цьому полі невалідні дані. В моєму випадку інструмент зміг визначити тип поля First_name та email. В полі First_name він хибно позначив велику кількість даних як не валідні за вбудованим патерном інструменту. Щодо поля email, то він правильно визначив поле, в яку я штучно ввела помилку (рис.3.17).

ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Services	Tariff plan	Account Status
85	Earvin	Majury	9/16/1992	94 Swallow Street	+7 500 212 4355	emajury2c/squidoo.com	100 Mbps	Internet	Gold Plan	Active

Рис.3.17 Виявлення помилки в полі email

Також по профілю даних можна помітити, що інструмент вдало визначив, що в полі Account status та Services присутні значення null, тобто, ці значення пусті(рис.3.18 та 3.19).

ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Services	Tariff plan	Account Status
18	Husein	Pellett	2/16/1969	386 Dryden Trail	+33 520 525 5592	hpelletth@marketwatch.com	1 Gbp	Internet	Basic Plan	

Рис.3.18 Пусте значення поля Account status

ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Services	Tariff plan	Account Status
54	Cecelia	Trewett	6/14/1993	7640 Cambridge Point	+502 979 416 2212	ctrewett1h@mac.com	100 Mbps		Basic Plan	Disconnected

Рис.3.19 Пусте значення поля Services

Окрім цього, в згенерованому мною датасеті є поля, які мають значення з певного переліку можливих варіантів. У таких полях зручно перевіряти кількість унікальних даних, яку також можна побачити по профілю даних, а також перевірити, які унікальні значення присутні в даному полі. В моєму випадку таких полів декілька: Connection type, Services, Tariff plan, Account status. Переглянувши статистику по унікальним значенням в цих полях, можна побачити, що існують значення, які не представлені в списках можливих значень (рис.3.20 – 3.23).

Value	Frequency	Percent
Cable TV	354	35%
Internet	326	33%
Phone	318	32%
Smartphone	1	0%
Null	1	0%

Рис.3.20 Статистика унікальних значень поля Services

Value	Frequency	Percent
1 Gbp	371	37%
100 Mbps	321	32%
50 Mbps	306	31%
one Gbp	1	0%
one hundred Mbps	1	0%

Рис.3.21 Статистика унікальних значень поля Connection Type

Value	Frequency	Percent
Gold Plan	345	34%
Basic Plan	327	33%
Premium Package	327	33%
Minimum Plan	1	0%

Рис.3.22 Статистика унікальних значень поля Tariff Plan

Value	Frequency	Percent
Disconnected	364	36%
Blocked	345	34%
Active	288	29%
Null	1	0%
Активний	1	0%
Блокировка	1	0%

Рис.3.23 Статистика унікальних значень поля Account Status

Після детального аналізу даних, можна переходити до наступного етапу. На етапі очищення та стандартизації даних, можна змінювати регістр полів за вимогами (рис.3.24). В моєму випадку поля First_name та Last_name повинні

відповідати певному регістру, який я і вказую для даних полів. Також на даному етапі можна використати власні патерни, для того щоб зпарсити значення рядка (використати це поле для фільтрації значень не вдалось).

Field Name	Type	Copy Field	Reverse Case	UPPER CASE	lower case	Proper Case	Remove Non-printable Characters	Replacement For Non-printable Characters	Replacement for Empty Values	Remove Leading Spaces	Remove Trailing Spaces	Characters To Remove	Characters To Replace	Remove Spaces	Remove Letters	Remove Numbers	Replace Zeros With O Letters	Replace O Letters With Zeros	Pattern Builder	Word Split
ID	-	☐	☐	☐	☐	☐	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
First_name	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Last_name	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Date of birth	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Address	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Phone Number	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Email	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Connection type	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Services	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Tariff plan	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒
Account Status	-	☐	☐	☐	☐	☒	☒	☒		☒	☒			☒	☐	☐	☐	☐	☒	☒

Рис.3.24 Вікно очищення та стандартизації даних

Також в інструментів можна використовувати злиття даних (рис.3.25) і маючи словник із можливими значеннями для певного поля, цю функцію можна використати для того, щоб відсіяти рядки, в яких значення поля не присутнє в словнику. Такий, свого роду костиль, я можу використати для полів Connection type, Services, Tariff plan, Account status, але варто зауважити, що через функціональні обмеження злиття в такому застосуванні може бути застосоване тільки для одного поля, що значно обмежує можливості.

Suggest Mapping		Split Horizontal		Match		Add Definition		Remove Definition		Remove Match Criteria		Max matches per result group: 0		Similar records in groups		Autogenerate report	
telecommunication06.csv		dictionary.csv		Include		Definition 1						Max matches per record: 500		Advanced options			
ID					☒												
First_name					☒												
Last_name					☒												
Date of birth					☒												
Address					☒												
Phone Number					☒												
Email					☒												
Connection type	Connect type				☒												
Services	Services				☒												
Tariff plan	Tariff plan				☒												
Account Status	Account Status				☒												

Рис.3.25 Вікно злиття даних

І як результат, дана утиліта змогла поправити регістр поля First_name в рядку з ID 7, а також за допомогою цієї утиліти вдалось відсіяти рядки 641 та 509, в яких були присутні значення поля Connection_type не зі списку. Також за допомогою цієї утиліти можна було виявити наявність рядків зі значенням

певних полів null, але відсіяти так поля не можна, і також по профілю можна було визначити, що в деяких полях існуються нестандартні значення.

Тому як результат, даний інструмент зміг виправити 3 помилки з 14, а також дозволив вручну виявити вади в деяких рядках у значенні поля Email, рядки із значенням null в деяких полях, а також нестандартні значення в полях, значення яких є елементами якогось списку, хоча якщо поле має багато унікальних значень, то таким методом визначити ваду не вийде. Тобто по суті він допоміг вручну виявити ще 7 помилок. Не виявленими залишились лише рядки з ID 68 в якому порушено атрибут достовірності - в полі Date_of_birth присутня адреса, яка ще не настала, 161 в якому порушується атрибут актуальності - різниця між актуальною датою та Date_of_birth більше ніж 110 років, тому в моєму випадку дані дані не актуальні, 343 та 438 в яких порушений атрибут відповідності - поле Phone number має номери телефону із більшою кількістю символів та меншою, ніж можливі (дозволено від 15 до 17 символів враховуючи '+' та пробіли).

7	Lacy	Hymus	1/30/2018	90272 Grasskamp Point	+86 930 740 0467	lhymus6@t.co	1 Gbp	Internet	Gold Plan	Blocked
9	Marcia	Downgate	1/17/1942	9768 Thackeray Center	+63 201 430 6893	mdowngate8@toplist.cz	1 Gbp	Smartphone	Basic Plan	Blocked
18	Husein	Pellet	2/16/1969	386 Dryden Trail	+33 520 525 5592	hpelletth@marketwatch.com	1 Gbp	Internet	Basic Plan	
39	Celene	Chismon	7/5/1962	77 Doe Crossing Place	+86 625 426 3921	cchismon12@macromedia.com	1 Gbp	Cable TV	Premium Package	Активний
54	Cecelia	Trewett	6/14/1993	7640 Cambridge Point	+502 979 416 2212	ctrewett1h@mac.com	100 Mbps		Basic Plan	Disconnected
68	Lindsay	Hedin	07/05/2050	74 Boyd Avenue	+62 732 280 3655	lhedin1v@dion.ne.jp	100 Mbps	Internet	Premium Package	Active
85	Earvin	Majury	9/16/1992	94 Swallow Street	+7 500 212 4355	emajury2c@squidoo.com	100 Mbps	Internet	Gold Plan	Active
161	Bastien	Paulet	01/06/1900	3 Toban Parkway	+86 647 181 0311	bpaulett4g@skype.com	50 Mbps	Phone	Basic Plan	Disconnected
164	Candide	Martschke	9/16/2014	265 Tennyson Alley	+62 291 350 7319	cmartschke4@amazon.co.jp	50 Mbps	Phone	Premium Package	Блокировка
293	Cointon	Ber	7/29/1972	7 Hanson Street	+93 406 959 8434	cber84@goo.gl	50 Mbps	Internet	Minimum Plan	Blocked
343	Kora	Merioth	6/13/1952	74955 Magdeline Circle	+502 914	kmerioth9@imgur.com	50 Mbps	Cable TV	Basic Plan	Disconnected
438	Courtney	Anyene	9/27/1994	3 Rusk Terrace	+375 684 228 1421 000	canyenec5@redcross.org	50 Mbps	Cable TV	Premium Package	Disconnected

Рис.3.26 Перевірка наявності помилок у вихідних даних

Цей інструмент є популярним на просторах інтернету як інструмент валідації даних для малого бізнесу. Інструмент Data Ladder дуже простий у використанні, але через таку простоту користувач платить функціональністю. Даний інструмент якісно виділився в профілюванні даних, так як він може автоматично розпізнавати тип поля по його даних, після чого перевіряти значення в цьому полі по власним словникам. Також інструмент має хороший функціонал з пошуку дублікатів у датасетах. Хоча простота інструменту позитивно впливає на користувацький досвід, через цю ж простоту даного інструменту, в ньому немає великої кількості тих функцій, які присутні в інших інструментах, наприклад створення персональних потоків завдань, в яким проводиться обробка даних, відсутні правила специфікацій, перевірка даних по певних умовах і ще купа різних функцій. Також під час роботи з інструментом виникали помилки при відкритті проекту, яка виникала випадковим чином. В цілому хоча даний інструмент дуже простий та зрозумілий у використанні, але його на мою думку може бути достатньо для одноразової перевірки певного датасету вручну, але не більше, так як відсутня будь-яка автоматизація перевірки, профілювання та навіть очищення даних. Результатом тестування інструменту на власному датасеті із штучно введеними помилками є 1 виправлений рядок із 14; 2 відкинуті рядки, з 14; а також 7 виявлених рядків з помилками за допомогою функцій інструменту із 14.

Інструмент Alteryx також є одним із популярних інструментів для валідації даних. З точки зору пробного доступу він якісно виділяється серед інших аналогів - для науковців та студентів він надає доступ на 1 рік використання. Також він має як хмарний інструмент, так і додаток для локального використання. На жаль, доступ до хмарного інструменту не вдалось отримати, так як доступ до нього надається по невідомим критеріям, тому протестовано лише додаток.

Далі зображено вигляд основного вікна програми (рис.3.27), видно, що вікно даного інструменту поділене на три області: область із функціональними

блоками, які поділені на категорії, область зліва, яка демонструє параметри налаштування вибраних блоків та справа - робоча область, куди можна перетягувати блоки і будувати функціональну схему, нижче робочої області розташоване поле результатів, де показана інформація про виконання і вивід тестових даних.

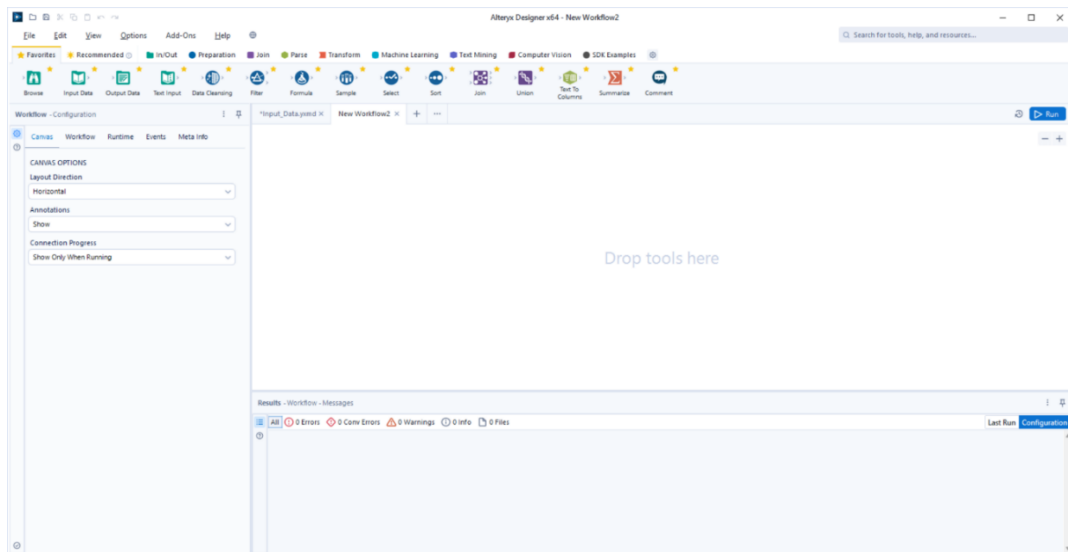


Рис.3.27 Основне вікно програми

Далі показано налаштування блоку Input Data (3.28). Даний блок потрібен для завантаження даних з певного джерела. Для того, щоб обрати джерело завантаження даних потрібно натиснути кнопку Set up a Connection.

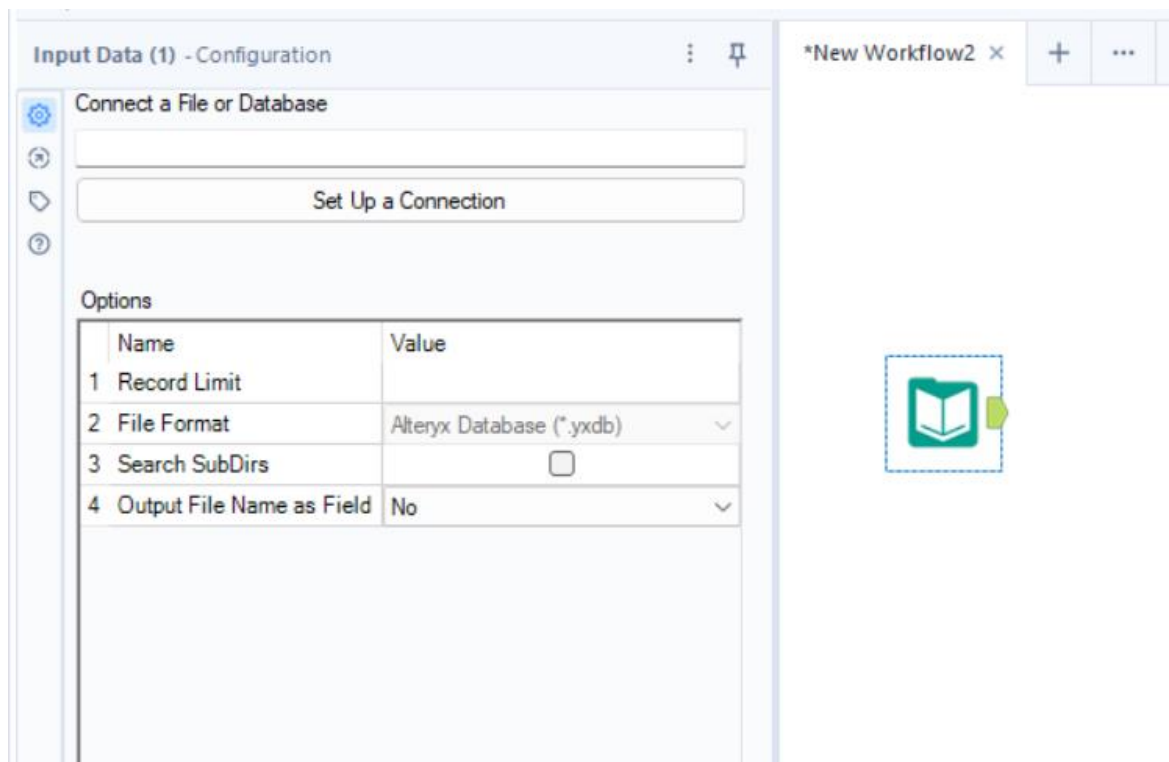


Рис.3.28 Налаштування блоку Input Data

Після натискання кнопки відкривається список доступних джерел. На рисунку 3.29 видно, що програма надає широкий вибір форматів фалів, з яких можна завантажити дані.

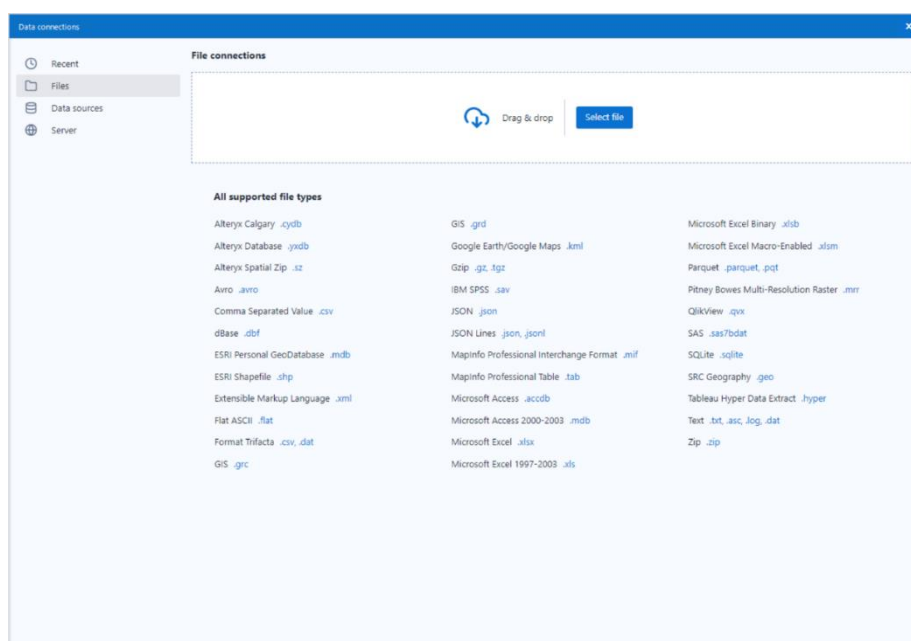


Рис.3.29 Список доступних форматів файлів для джерел даних

Також програма дозволяє працювати із великою кількістю підключень, включаючи системи управління базами даних, онлайн - сервіси та хмарні сховища даних (рис.3.30).

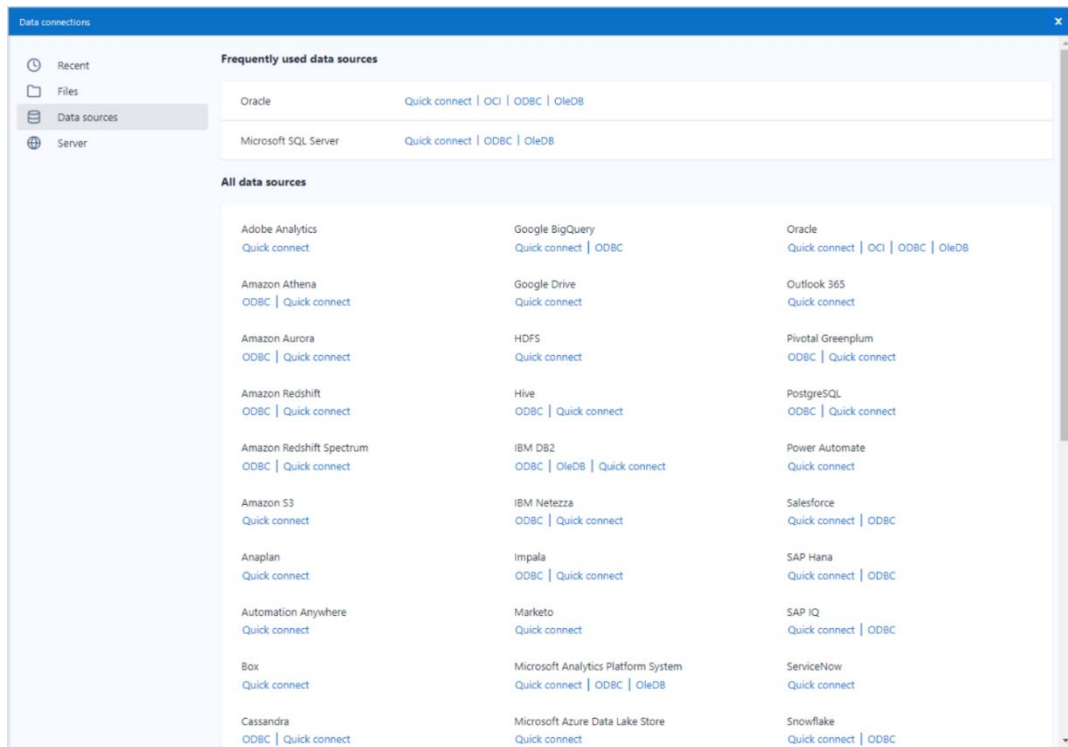


Рис.3.30 Список можливих підключень до зовнішніх джерел

Далі продемонстровано процес створення нового з'єднання із СУБД, на прикладі PostgreSQL(рис.3.31), для того, щоб створити нове з'єднання потрібно вказати назву джерела, вказану при налаштуванні драйвера, адресу сервера, на якому розміщена БД, порт та назву самої БД, до якої проводиться підключення.



Рис.3.31 Процес створення нового з'єднання із СУБД

Інструмент надає можливість використовувати різні методи для авторизації, я використала Username and password (рис.3.32).

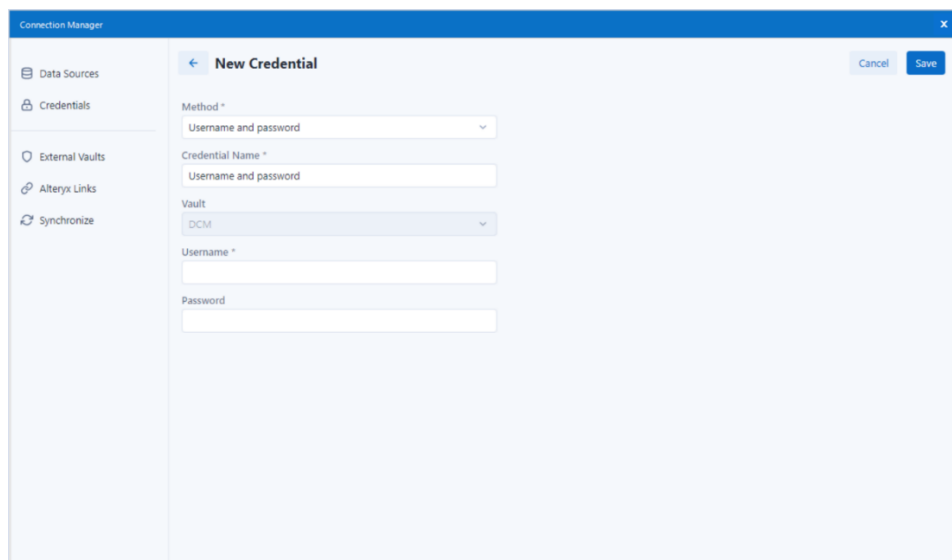


Рис.3.32 Вікно авторизації

Для того, щоб авторизуватись під час підключення до сервера потрібно з'єднати дані для авторизації із джерелом даних (рис.3.33).

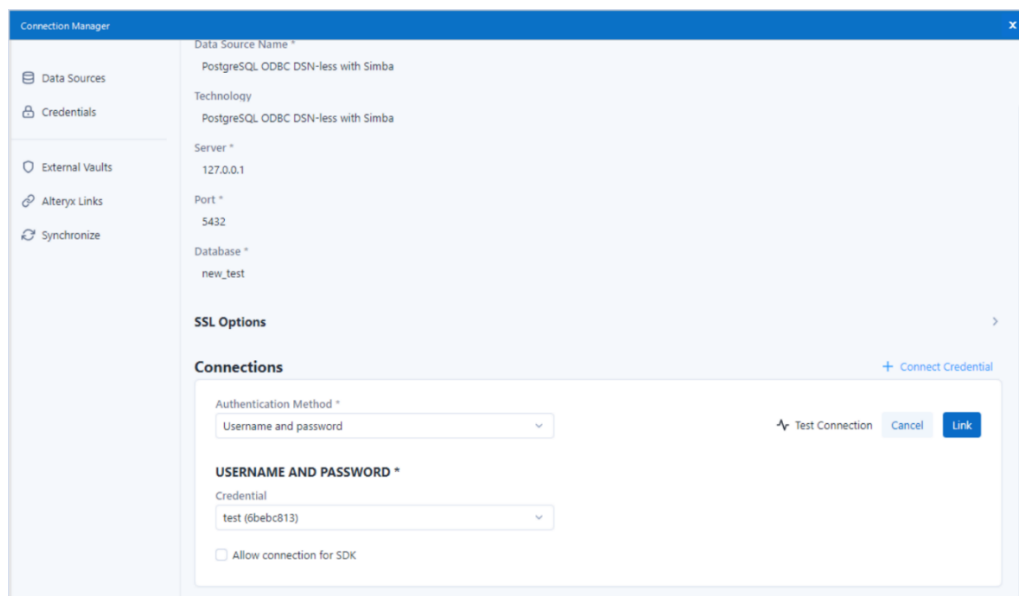


Рис.3.33 З'єднання даних для авторизації із джерелом даних

Коли обирається джерело для завантаження даних, то відкривається вікно із додатковими налаштуваннями для їх розпізнавання та форматування. На рисунку 3.34 зображено вікно, на якому можна попередньо переглянути коректність відображення вибраних даних.

Options

	Name	Value
1	Record Limit	
2	File Format	Comma Separated Va ▾
3	Search SubDirs	☐
4	Output File Name as Field	No ▾
5	Delimiters	.

Preview (first 100 records) Refresh

	ID	First_name	Last_name	Date of birth	Address
1	1	Nessa	Larkkem	10/25/1974	0 Tomscot Hill
2	2	Burke	Billham	5/4/2004	Null
3	3	Emmaline	Deinhardt	2/26/2000	[Null]
4	4	Marcello	Janz	[Null]	2075 Helena Avenue
5	5	[Null]	Testin	4/6/1940	18873 Sunbrook Alle
6	6	Micheil	[Null]	1/30/1942	[Null]
7	7	Nester	Smaleman	10/25/1949	36 Weeping Birch Pl
8	8	Wyatt	Benedict	9/2/1981	1 Dwight Center
9	9	Sarina	Brundle	6/18/1972	2833 Eastlawn Plaza
10	10	Julina	Montfort	[Null]	4708 Sage Center
11	11	Jeanna	Vaud	8/6/2000	3026 Summerview C
12	12	Bartram	Null	9/22/1977	0603 Schmedeman l
13	13	Orin	[Null]	6/20/2007	0 Longview Way
14	14	Earlie	Rubinfeld	[Null]	80347 Hanover Cros
15	15	Null	Coupe	4/25/1961	718 Mendota Avenue
16	16	Kane	Stathers	6/14/1982	7564 Carpenter Hill
17	17	Mari	Goldsack	[Null]	8 Dexter Park
18	18	Daisey	[Null]	7/14/1960	5154 Banding Street

Рис.3.34 Додаткові налаштування розпізнавання та форматування даних

Далі можна побачити загальний вивід блоку Browse. Даний блок використовується для перегляду даних без збереження в певне джерело. Знизу відображаються всі внесені дані, а зліва профілювання цих даних (рис.3.35).

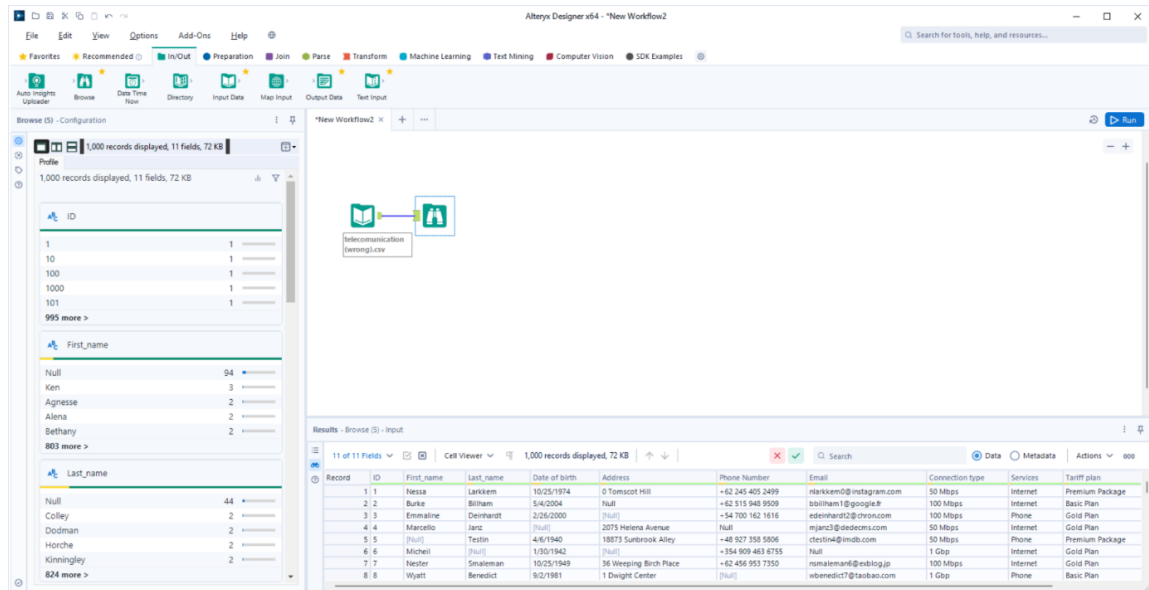


Рис.3.35 Вивід блоку Browse

Вигляд профілювання даних за допомогою блоку Browse, зелена лінія демонструє, яка частка даних валідна, і жовта - вказує на частку нульових значень в даному полі(рис.3.36).

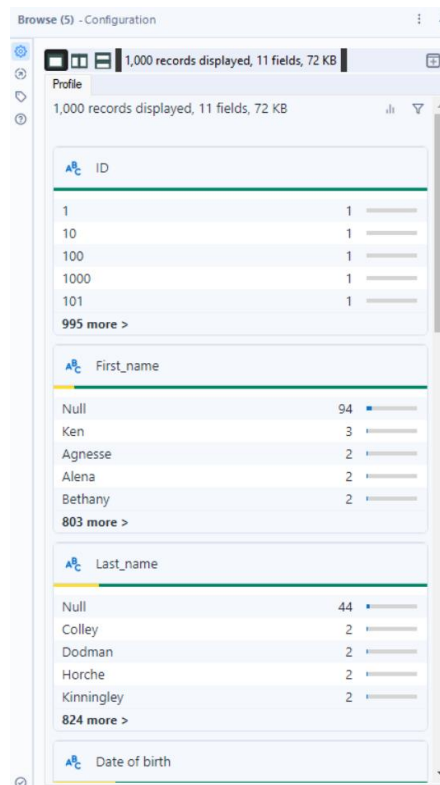


Рис.3.36 Профілювання даних за допомогою блоку Browse

Для кожного поля можна переглянути детальну статистику. Далі зображена статистика на прикладі поля First_name (рис.3.37). Можна переглянути детальний звіт про те, скільки відсотків коректних даних, скільки унікальних, яка кількість нульових значень та пустих клітинок. Також відображена статистика про довжину значень, яке мінімальне, максимальне і середнє значення довжини поля First_name.

Дана статистика важлива для того, щоб оцінити якість внесених даних і необхідність їх виправлення.

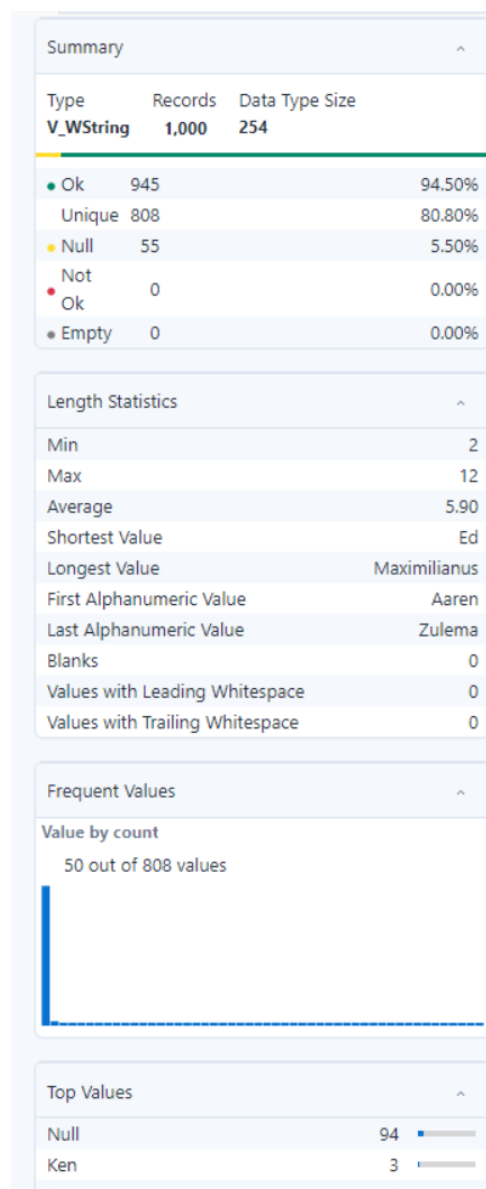


Рис.3.37 Статистика поля First_name

Також в інструменті наявний блок Sort, який дозволяє сортувати дані, name - вказує на ім'я поля за яким відбувається сортування даних, order - в якому порядку відбувається сортування - зростання або спадання (рис.3.38).

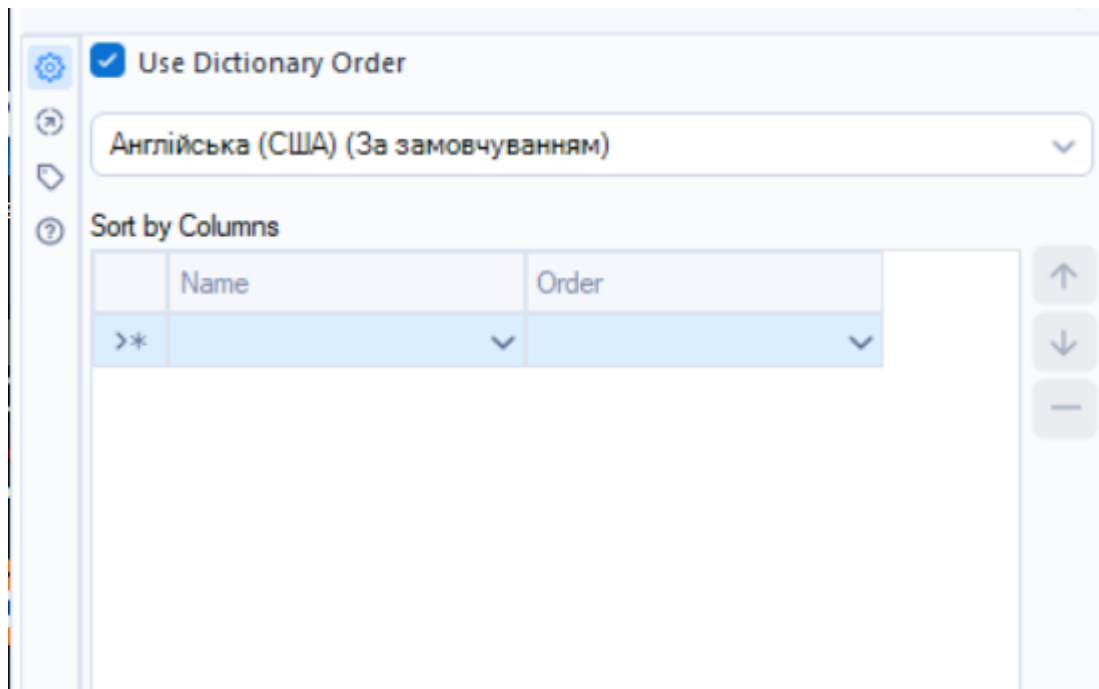


Рис.3.38 Вивід блоку Sort

Також є блок Data Cleansing, який призначений для очищення даних: видалення пустих рядків, полів і колонок. Також дана функція надає можливість замінювати нульові значення на 0 або Blanks, в залежності від того який це тип даних. Також можна видаляти пробіли, дублікати пробілів, розриви рядків, букви, цифри і пунктуаційні знаки, також можна редагувати регістри (рис.3.39).

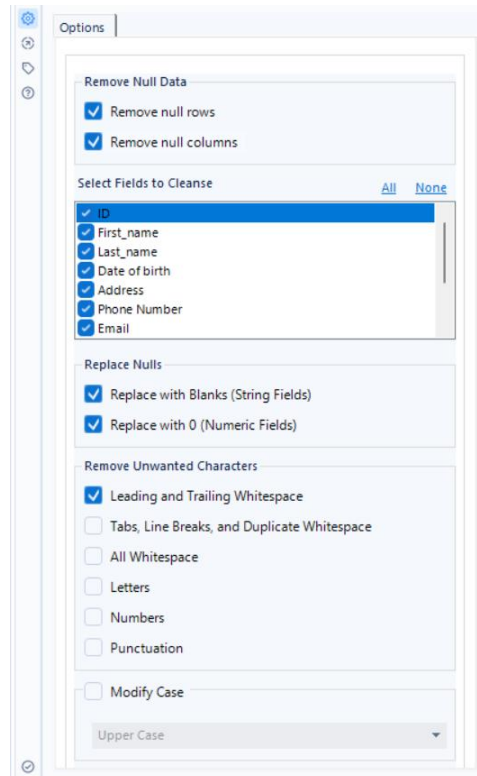


Рис.3.39 Вивід блоку Data Cleansing

Наступним корисним блоком є Filter. Налаштування даного блоку допомагає перевіряти дані на виконання певної умови, завдяки чому можна відокремити підходящі дані від тих, які не підходять під умову (рис. 3.40).

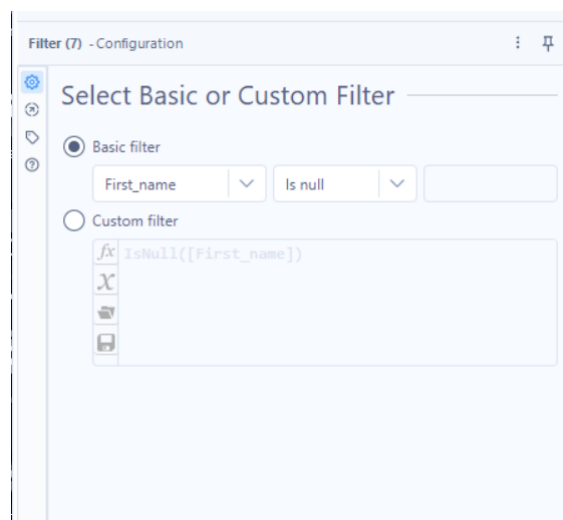


Рис.3.40 Налаштування блоку Filter

Далі зображена перевірка на наявність пустих полів і на рівність поля значенню Null за допомогою умови (рис.3.41). Якщо поле містить пусті рядки, пусті значення або Null, то це неякісні дані.

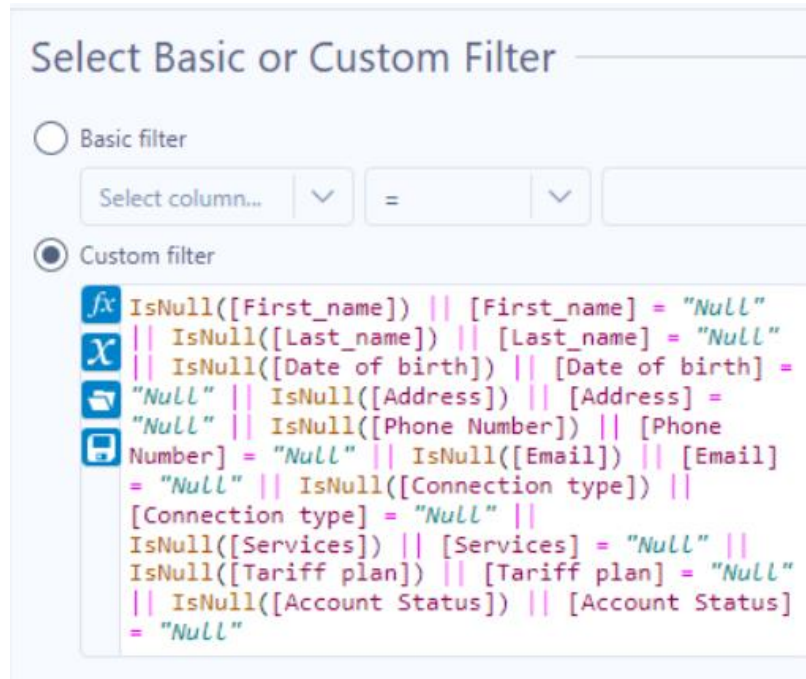


Рис.3.41 Фільтри для відокремлення даних

Далі зображено вигляд робочого потоку для очищення даних від пустих значень або Null (рис.3.42).

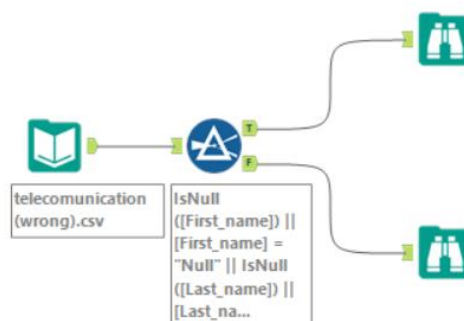


Рис.3.42 Блок-схема для очищення даних

Так як дублікати - це неякісні дані, програма має блок Unique призначений для того, щоб обирати унікальні рядки за певними полями, тобто відділяти дублікати із датасету (рис.3.43).

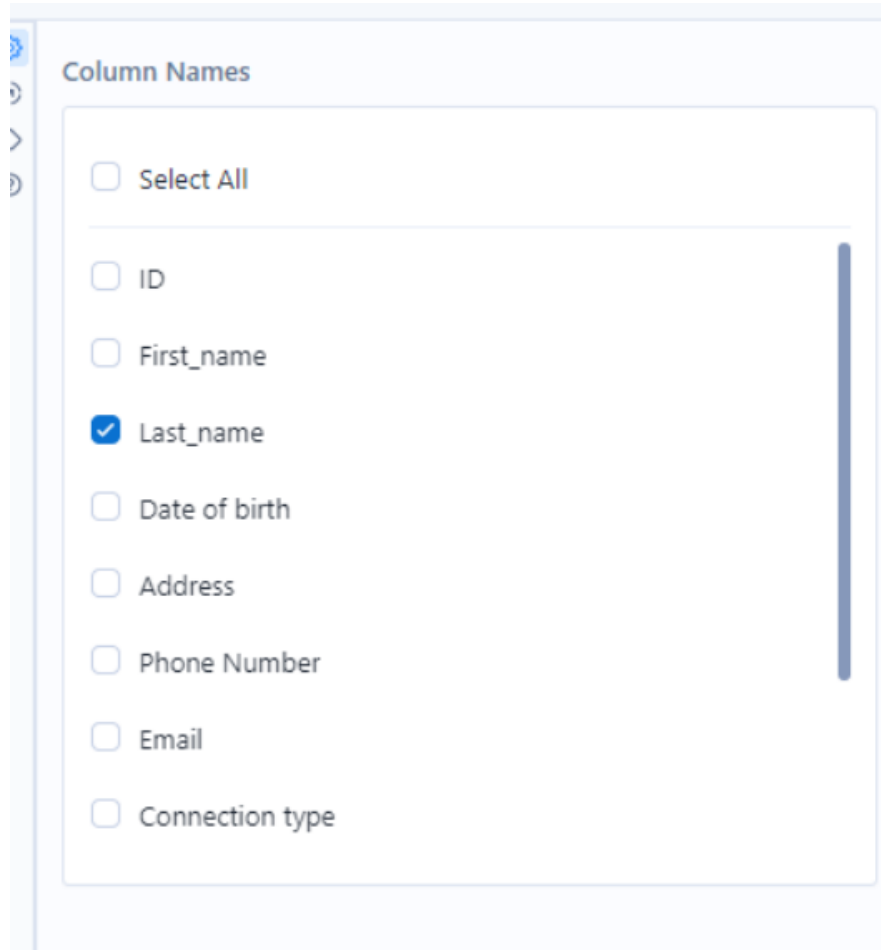


Рис.3.43 Налаштування блоку Unique

Блок Output Data - призначений для того, щоб зберігати оброблені дані до певного джерела(рис.3.44)

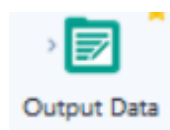


Рис.3.44 Блок Output Data

Інструмент дозволяє в налаштуваннях цього блоку обрати різні формати файлів для збереження даних (рис. 3.45).

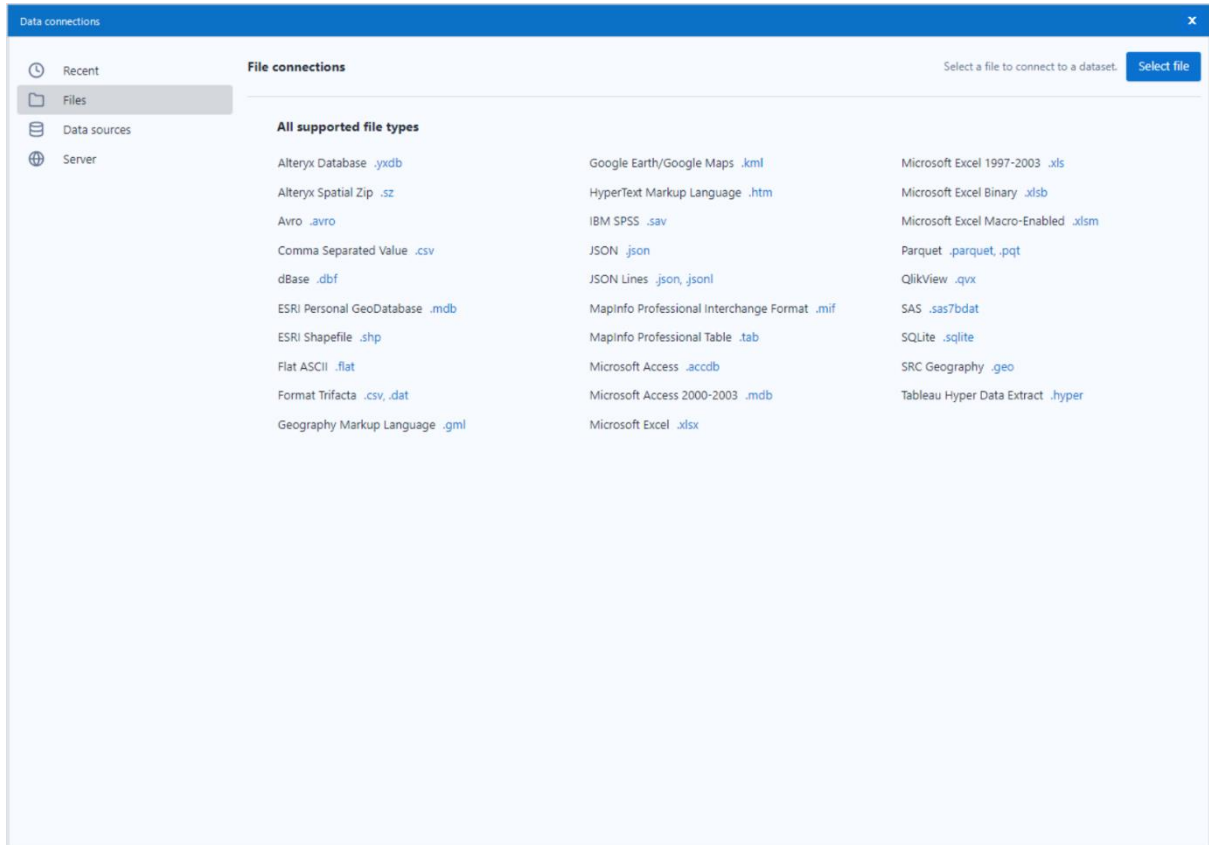


Рис.3.45 Налаштування блоку Output Data

Також можливі способи збереження даних не у вигляді файлу. Інструмент дозволяє зберігати дані у великій кількості популярних СУБД, а також в хмарних сервісах (рис.3.46).

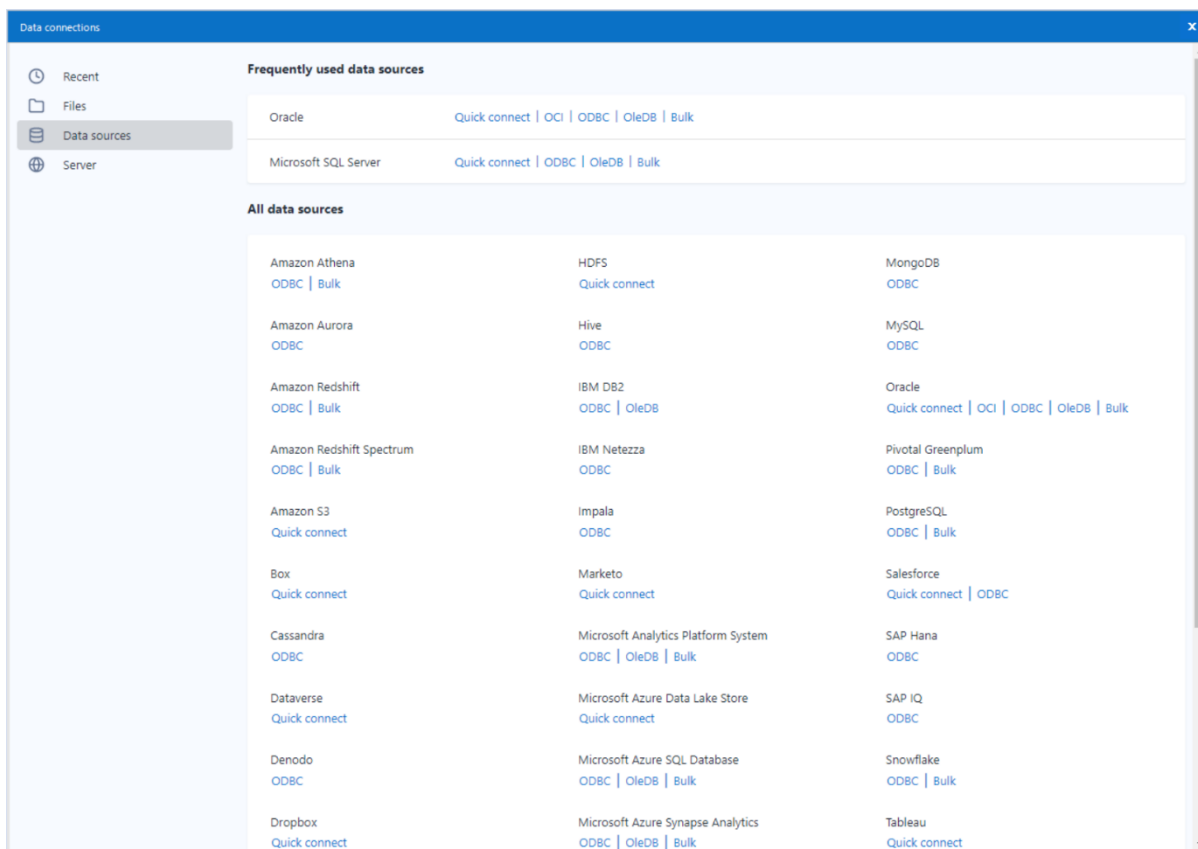


Рис.3.46 Варіанти підключень до СУБД або хмарних сховищ

Після того як обрано джерело збереження файлу, можна обрати додаткові налаштування, такі як формат, вид роздільника (двокрапка, кома чи крапка з комою) для файлу формату csv, можна вибрати спосіб кодування (рис.3.47).

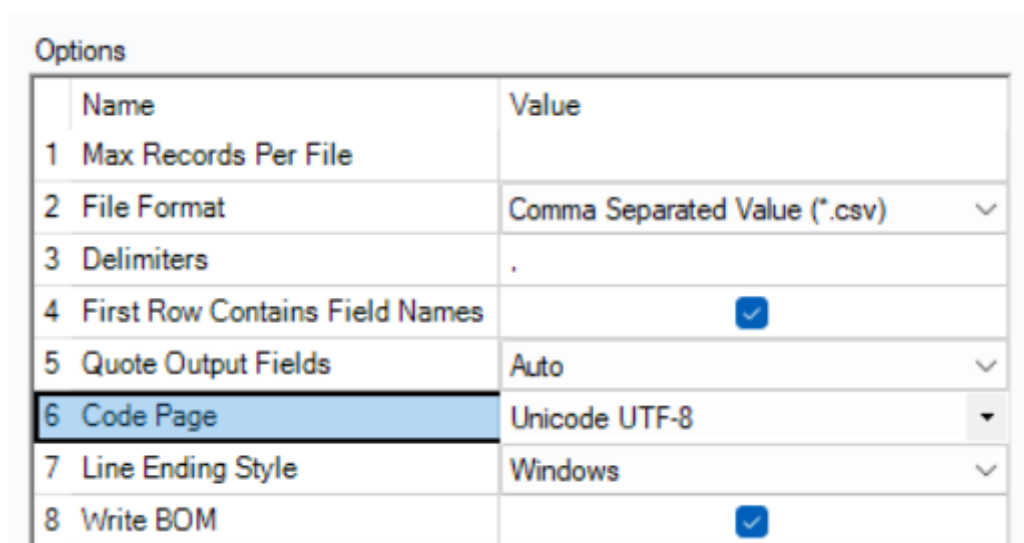


Рис.3.47 Додаткові налаштування вибраного джерела

Такий вигляд має блок ра допомогою якої із джерела можна відфільтрувати рядки із пустими полями, а також із полями які містять значення Null, також видаляти дублікати. На виході ми маємо збережені дані, які не будуть мати пустих полів та значень, а також дані не будуть мати дублікатів (рис. 3.48).

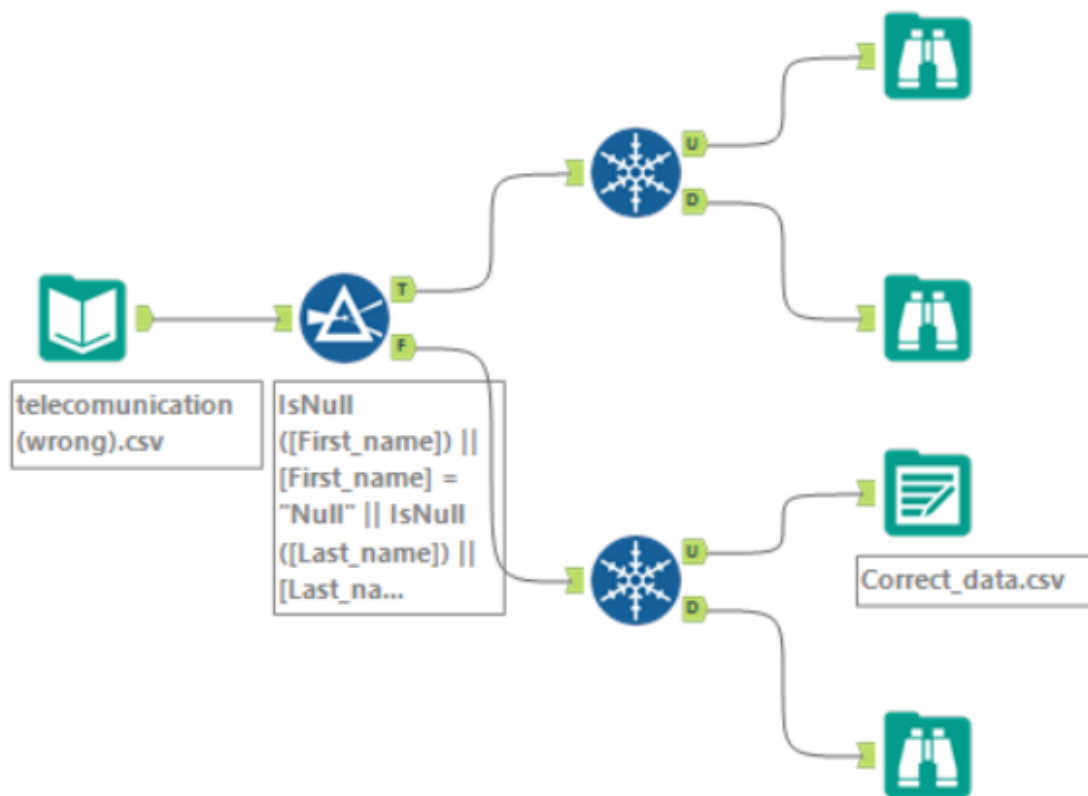


Рис.3.48 Блок-схема для відбору неякісних даних

Тестування на власному датасеті проводилось також починаючи із профілювання. Завдяки цьому кроку можна помітити, що в деяких полях є клітинки зі значенням Null (рис.3.49-3.50). Також можна побачити розподіл частки заповненості поля конкретними значеннями.

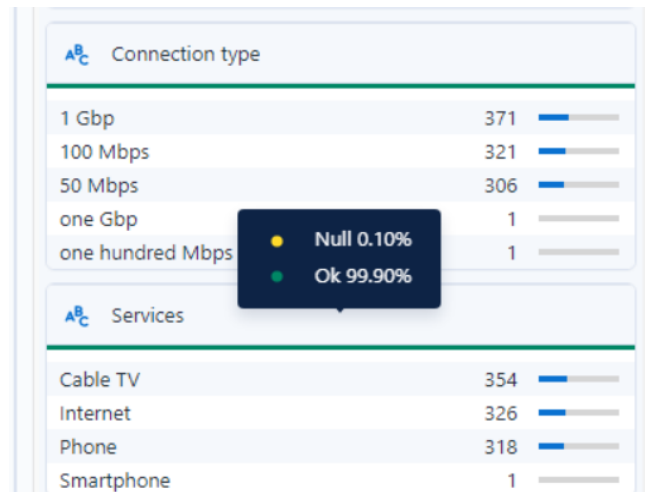


Рис.3.49 Нульові значення в полі Services

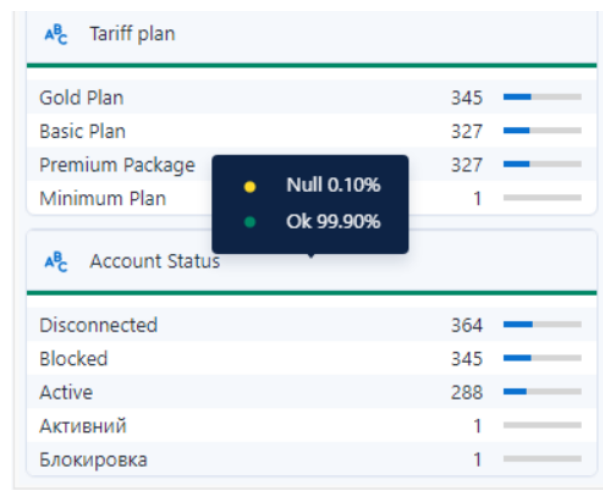


Рис.3.50 Нульові значення в полі Account Status

Після кроку із профілюванням я приступила до побудови робочого потоку для очищення даних від всіх помилок, які були введені в мій датасет. Фінальна схема робочого потоку зображена нижче (рис.3.51).



Рис.3.53 Ділянка очищення значень

Наступною ділянкою є ділянка, в якій перевірялись значення полів Connection type, Services, Tariff plan та Account status на присутність в словнику Tariff plan dict, в який занесено по стовпцям можливі значення для кожного поля (рис.3.54).

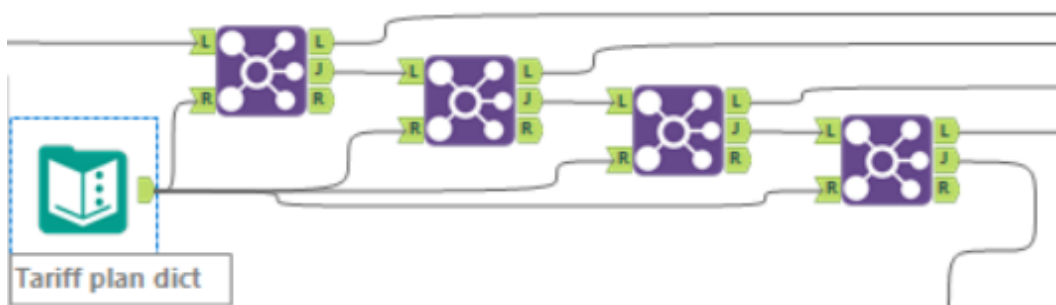


Рис.3.54 Ділянка перевірки значень полів Connection type, Services, Tariff plan та Account status

Наступним блоком є фільтр, в якому перевірялась довжина значень поля Phone Number на входження в границі - від 15 до 17 символів включно, враховуючи знак '+' та пробіли (рис.3.55).

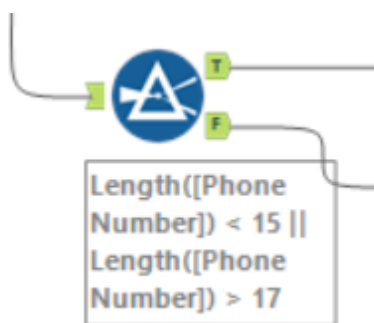


Рис.3.55 Фільтр перевірки значень поля Phone Number

Наступним блоком є також фільтр, в якому за допомогою regex патерну перевірялись значення поля Email (рис.3.56).

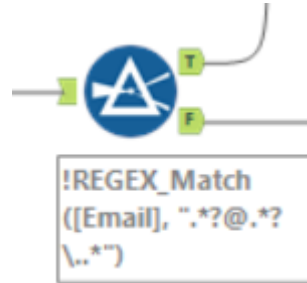


Рис.3.56 Фільтр перевірки значень поля Email

І останнім фільтром присутнім в схемі робочого потоку проводиться перевірка актуальності даних (рис.3.57). В моєму випадку я окреслила, що актуальними даними будуть ті, в яких вік користувача не перевищує 110 років і також достовірними даними будуть ті, в яких вік користувача більше за 5 років.

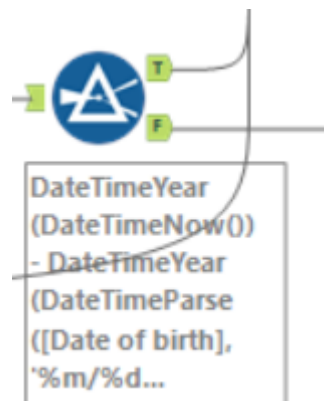


Рис.3.57 Фільтр перевірки актуальності даних

І фінальними частинами є дві кінцеві частини робочого процесу. На першій збираються невалідні дані після чого сортуються (рис.3.58).

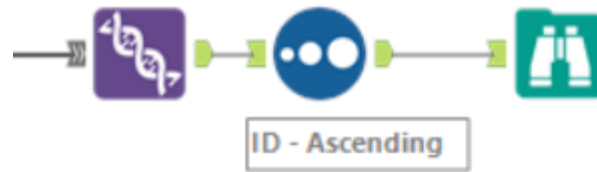


Рис.3.58 Об'єднання і сортування невалідних даних

І як результат ми можемо побачити відокремлені невалідні дані (рис.3.59), які можна зберегти в окремий файл або базу даних, задля подальшої обробки, або просто нічого з ними не робити.

ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Services	Tariff plan	Account Status
9	Marcia	Downgate	1/17/1942	9768 Thackeray Center	+63 201 430 6893	Mdowngate8@Toplist.Cz	1 Gbp	Smartphone	Basic Plan	Blocked
18	Husein	Pellet	2/16/1969	386 Dryden Trail	+33 520 525 5592	hpelletth@marketwatch.com	1 Gbp	Internet	Basic Plan	[Null]
39	Celene	Chismon	7/5/1962	77 Doe Crossing Place	+86 625 426 3921	Cchismon12@Macromedia.Com	1 Gbp	Cable Tv	Premium Package	Активний
54	Cecelia	Trewett	6/14/1993	7640 Cambridge Point	+502 979 416 2212	ctrewett1h@mac.com	100 Mbps	[Null]	Basic Plan	Disconnected
68	Lindsay	Hedin	07/05/2050	74 Boyd Avenue	+62 732 280 3655	Lhedinv@Dion.Ne.Jp	100 Mbps	Internet	Premium Package	Active
85	Earvin	Majury	9/16/1992	94 Swallow Street	+7 500 212 4355	Emajury2c@Squidoo.Com	100 Mbps	Internet	Gold Plan	Active
161	Bastien	Pauleit	01/06/1900	3 Toban Parkway	+86 647 181 0311	Bpauleit4g@Skype.Com	50 Mbps	Phone	Basic Plan	Disconnected
164	Candide	Martschke	9/16/2014	265 Tennyson Alley	+62 291 350 7319	Cmartschke4j@Amazon.Co.Jp	50 Mbps	Phone	Premium Package	Блокировка
293	Cointon	Ber	7/29/1972	7 Hanson Street	+93 406 959 8434	Cber84@Goo.Gl	50 Mbps	Internet	Minimum Plan	Blocked
343	Kora	Merioth	6/13/1952	74955 Magdeline Circle	+502 914	Kmerioth9i@Imgur.Com	50 Mbps	Cable Tv	Basic Plan	Disconnected
438	Courtney	Anyene	9/27/1994	3 Rusk Terrace	+375 684 228 1421 000	Canyenec5@Redcross.Org	50 Mbps	Cable Tv	Premium Package	Disconnected
509	Sibyl	Pashby	10/29/2007	9975 Mallard Street	+380 221 323 6833	Spashbye4@Reverbnation.Com	One Gbp	Phone	Basic Plan	Blocked
641	Willie	Braghini	3/23/1988	79098 Hermina Point	+380 880 886 2796	Wbraghinihs@Booking.Com	One Hundred Mbps	Internet	Gold Plan	Active

Рис.3.59 Вивід невалідних даних

Другою кінцевою частиною робочого процесу є частина, в якій сортуються та виводяться валідні дані (рис.3.60).

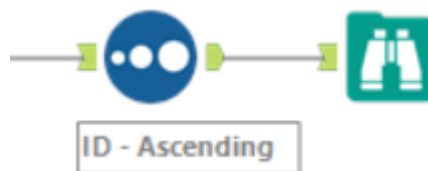


Рис.3.60 Сортування та вивід валідних даних

Після виводу, валідні дані можна переглянути (рис.3.61), або знову ж таки зберегти в окремий файл або базу даних.

ID	First_name	Last_name	Date of birth	Address	Phone Number	Email	Connection type	Services	Tariff plan	Account Status
1	Rita	Reye	12/19/1943	42 Roth Center	+81 231 330 5460	Rreye0@Zdnet.Com	1 Gbp	Phone	Basic Plan	Active
2	Dag	Abelson	11/20/1967	4 Debs Court	+974 558 615 7915	Dabelson1@Goodreads.Com	50 Mbps	Phone	Premium Package	Blocked
3	Merrily	Cain	12/28/1945	77713 Glendale Center	+55 870 951 2916	Mcain2@Paypal.Com	1 Gbp	Phone	Premium Package	Blocked
4	Madelyn	Lorinez	9/19/2016	7186 Norway Maple Place	+62 280 472 0673	Mlorinez3@Ning.Com	50 Mbps	Cable Tv	Basic Plan	Disconnected
5	Leroi	Melan	7/5/2016	04 Namekagon Street	+86 809 382 8164	Lmelan4@1688.Com	100 Mbps	Cable Tv	Gold Plan	Disconnected
6	Andie	Becerro	8/28/1988	39 Calypso Pass	+33 115 805 1749	Abecerro5@Howstuffworks.Com	100 Mbps	Cable Tv	Gold Plan	Blocked
7	Lacy	Hymus	1/30/2018	90272 Grasskamp Point	+86 930 740 0467	Lhymus6@T.Co	1 Gbp	Internet	Gold Plan	Blocked
8	Deva	Caldera	2/10/1999	7 Bluestem Lane	+98 230 726 9651	Dcaldera7@Webs.Com	100 Mbps	Internet	Basic Plan	Active
10	Augustus	Callam	9/21/1966	3 Rockefeller Center	+62 975 296 6695	Acallam9@Reddit.Com	1 Gbp	Phone	Basic Plan	Active
11	Stavro	De Vaux	8/5/1945	4356 Cody Alley	+993 865 715 8575	Sdevauxa@Census.Gov	1 Gbp	Cable Tv	Basic Plan	Active
12	Virgie	Bakey	10/18/1983	2559 6th Circle	+48 998 111 4907	Vbakeyb@Independent.Co.Uk	1 Gbp	Internet	Basic Plan	Active
13	Phaedra	Lippini	1/11/1988	07427 Clarendon Circle	+62 984 334 5720	Plippinic@Senate.Gov	50 Mbps	Cable Tv	Basic Plan	Blocked
14	Kaleena	Battlestone	11/3/1967	6 Kedzie Circle	+86 893 368 8792	Kbattlestone@Homestead.Com	1 Gbp	Cable Tv	Gold Plan	Disconnected
15	Heath	Cuthbertson	11/26/1943	36 Butterfield Road	+380 508 317 0245	Hcuthbertson@Dyndns.Org	1 Gbp	Cable Tv	Basic Plan	Disconnected

Рис.3.61 Вивід валідних даних

Результатом тестування інструменту на прикладі мого датасету є обробка 14 помилок із 14, і одну з них, а саме неправильний регістр значення, було з легкістю виправлено, і такі дані не були відокремлені. Також щодо відокремлених даних, то маючи якісь визначені критерії для обробки помилок, в даному інструменті можна автоматизувати їхню обробку згідно визначеним критеріям.

Інструмент Alteryx показав себе з хорошої сторони при валідації даних. Даний інструмент має зручний інтерфейс, з яким не складно розібратись навіть тому, хто вперше стикається з інструментами такого типу, а також деякі інструменти валідації даних мають схожий інтерфейс, завдяки чому на даний інструмент буде легко перейти, якщо користувач працював із схожими інструментами. Alteryx має велику кількість потрібних функцій для валідації даних, таких як профілювання даних, їх очищення та трансформування, перевірки по правилу, виявлення дублікатів та їхнє видалення і купа інших. Також даний інструмент має високу швидкість відгуку та виконання тестів. При виконанні будь-яких тестів, виконання супроводжується інформуванням про теперішній крок, помилки та попередження при виконанні сценарію і так далі. Таке журналювання виконання сценаріїв є дуже зручним, так як користувач у випадку виникнення помилок під час виконання сценарію з легкістю зможе виявити, в чому саме проблема, і виправити неполадку. Також даний інструмент має велику кількість документації, яка допомагає при вивченні функціоналу інструмента. В разі виникнення проблеми, яку складно виправити, можна звернутись до служби підтримки, яка оперативно допоможе у вирішенні

проблеми. Після використання даного інструменту в мене залишились лише позитивні емоції, тому я рекомендую даний інструмент для ознайомлення тим, хто в пошуках зручного, красивого, а головне функціонального інструменту валідації даних. Результатом обробки даним інструментом власного датасету із штучно введеними помилками є оброблені 14 помилок з 14, одну з яких, а саме неправильний регістр значень, було з легкістю оброблено, через що й не відокремлено в окремий потік.

Наступним протестованим інструментом для валідації даних є Informatica Data Quality. Цей інструмент відрізняється від інших тим, що представлений він не додатком, а хмарним сервісом. Реєстрація в даному інструменті проста, пробний період триває 30 днів. Після реєстрації для роботи на сервісі в електронному листі приходить посилання на сайт, на якому буде відбуватись робота. Це обумовлено тим, що сервіс хоститься на серверах з різних регіонів світу, і для кожного регіону виділене окреме посилання.

Отож, після входу в особистий профіль на сайті інструменту користувача зустрічає сторінка із доступними сервісами (рис.3.62).

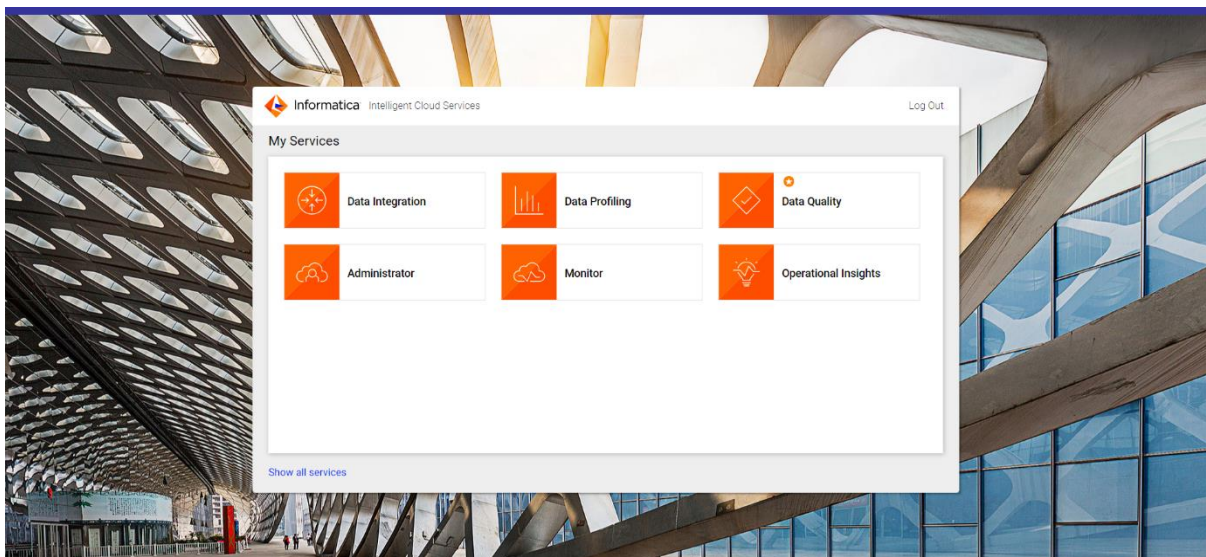


Рис.3.62 Список доступних сервісів

Кожен сервіс має свій функціонал, і загалом такий поділ функцій по сервісам є зручним, так як інтерфейс інструменту не перевантажений різноманітністю можливостей.

Сервіс Administrator призначений для керування налаштуваннями: налаштування персональних даних, керування дозволами, створення нових користувачів, так як є можливість використовувати інструмент командою, налаштування з'єднань, додавання наборів передналаштованих асетів і так далі(рис.3.63).

The screenshot displays the Informatica Administrator web interface. On the left is a navigation sidebar with icons for Organization, Licenses, SAML Setup, Metering, Users, Settings, User Groups, User Roles, Runtime Environ..., Connections, Add-On Connect..., Schedules, Add-On Bundles, Swagger Files, and Logs. The main content area is titled 'KPI' and 'Specify the organization details'. It contains an 'Organization Overview' form with two columns: 'General' and 'Address'. The 'General' column includes fields for 'Organization Name' (filled with 'KPI'), 'Organization ID' (filled with 'kFYzFxb8L4UIWizH-HyB4i'), 'Environment Type' (a dropdown menu currently showing 'Production'), 'Description' (a text area), and 'Number of Employees' (a dropdown menu currently showing 'Fewer than 10 employees'). The 'Address' column includes fields for 'Address 1', 'Address 2', 'Address 3', 'City', 'State', and 'Zip Code', all of which are currently empty. A 'Save' button is located in the top right corner of the form area.

Рис.3.63 Вікно сервісу Administrator

Сервіс Administrator

Корисним налаштуванням в даних сервісах є налаштування середовищ виконання. Даний інструмент не дозволяє виконувати аналіз даних на власних потужностях, і тому потрібно створювати власне середовище виконання (рис.3.64). Це можна зробити шляхом встановлення безпечного агента на локальний пристрій, такий як ПК, або шляхом встановлення цього агента на сервер. Агент підтримує операційні системи Linux та Windows, але явним недоліком є те, що він досі не підтримує Windows 11: він встановлюється, але

не дозволяє взаємодіяти з файловою системою пристрою, на який встановлений.

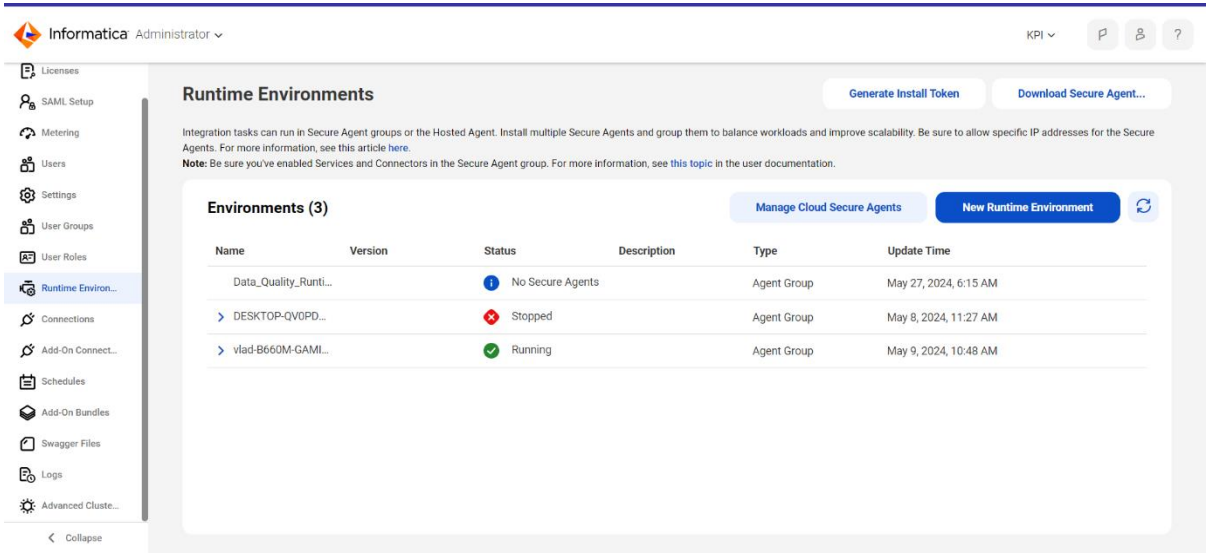


Рис.3.64 Існуючі середовища виконання, включаючи власноруч створене.

Наступним важливим кроком є налаштування підключень. На даному етапі можна створити підключення, з яких в майбутньому можна буде отримувати дані, а також в які можна буде зберігати дані (рис.3.65). Тут можна створити підключення з різними СУБД, підключення з директорією, з якої можна буде потім обирати файли, а також зберігати, підключення з хмарними сервісами і також з драйвером ODBC, завдяки якому можна створювати підключення з безліччю СУБД. Важливим також є той факт, що при підключенні до СУБД по налаштуванням які надає сервіс, функціонал буде відрізнятись від підключення до СУБД через ODBC.

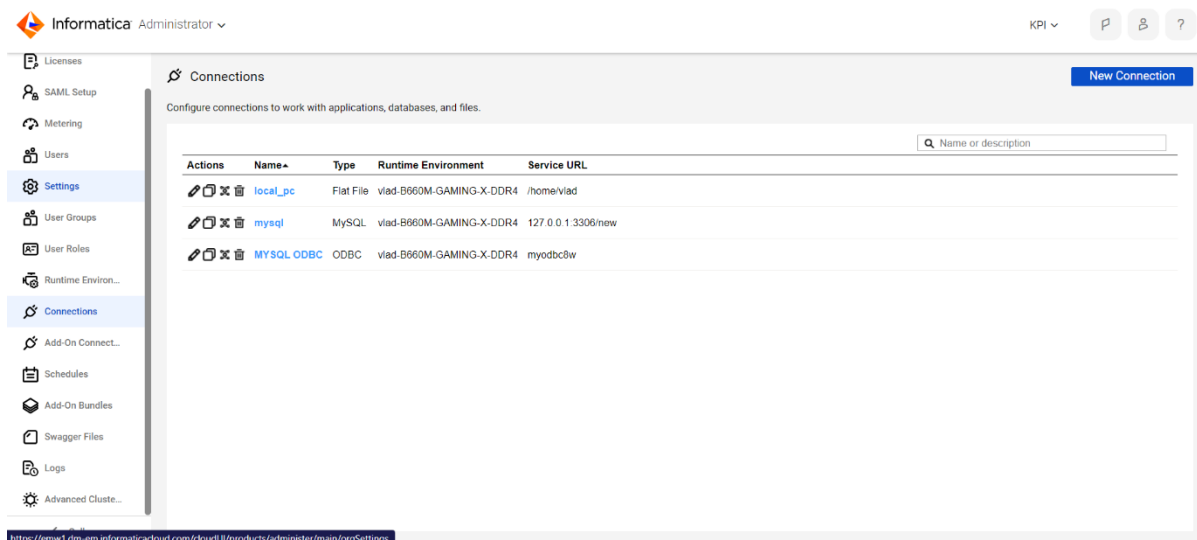


Рис.3.65 Налаштовані підключення

Після налаштувань середовищ виконання та підключень, що дозволяє взаємодіяти з сховищами даних та виконувати обробку та аналіз даних, можна переходити далі.

Наступним важливим сервісом є Data Profiling (рис.3.66).

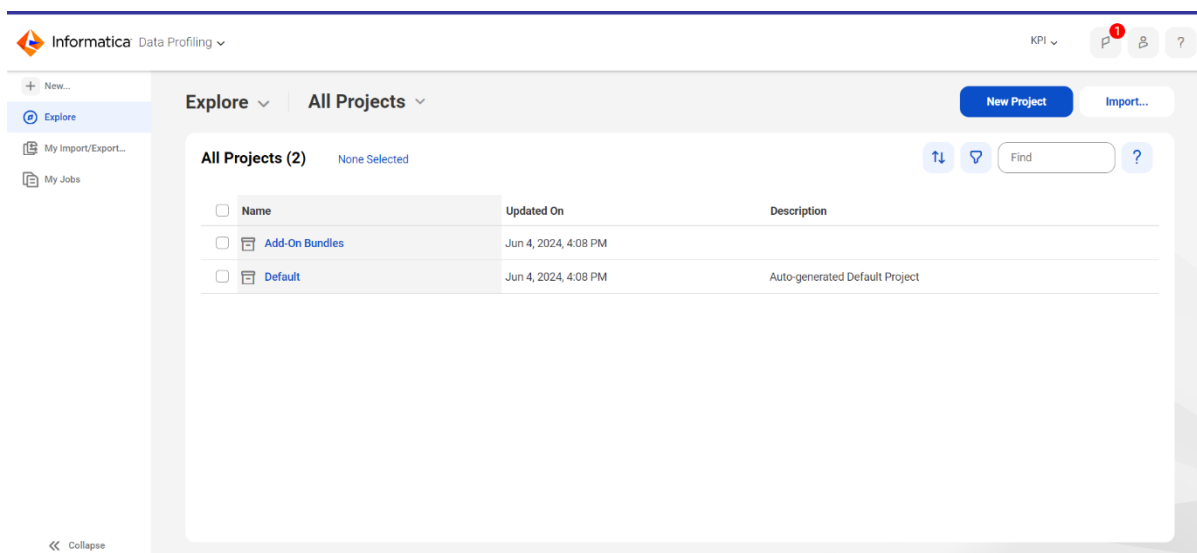


Рис.3.66 Головне вікно сервісу профілювання

В даному сервісі можна створити лише один об'єкт, а саме профілювання даних (рис.3.67).

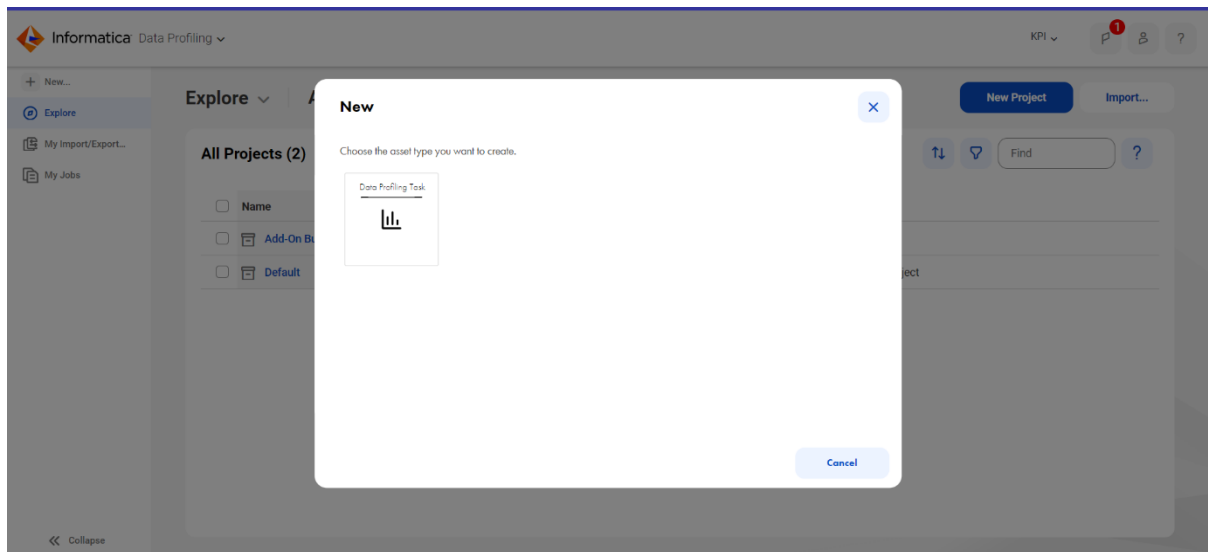


Рис.3.67 Можливі варіанти об'єктів у сервісі Data Profiling

Даний об'єкт, як і будь-який створюваний об'єкт, можна іменувати а також додати опис. Після цього потрібно обрати джерело даних(рис.3.68).

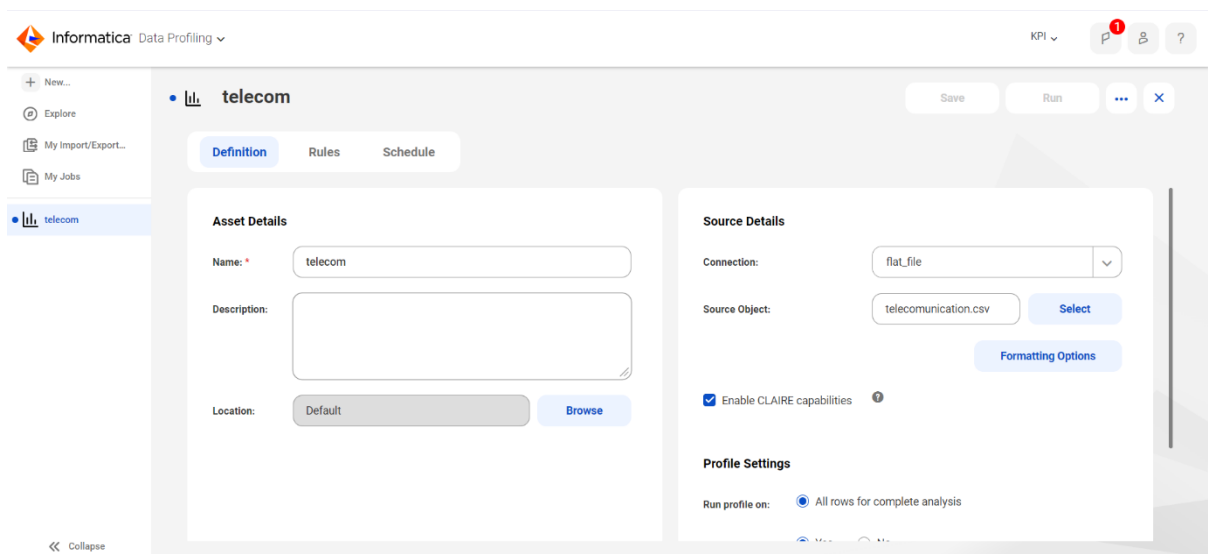


Рис.3.68 Налаштування профілювання

Після проведення налаштування, знизу можна побачити прев'ю як дані будуть підвантажені в сервіс (рис.3.69). Також там же можна обрати, дані з яких полів потрібно завантажувати, а з яких не потрібно.

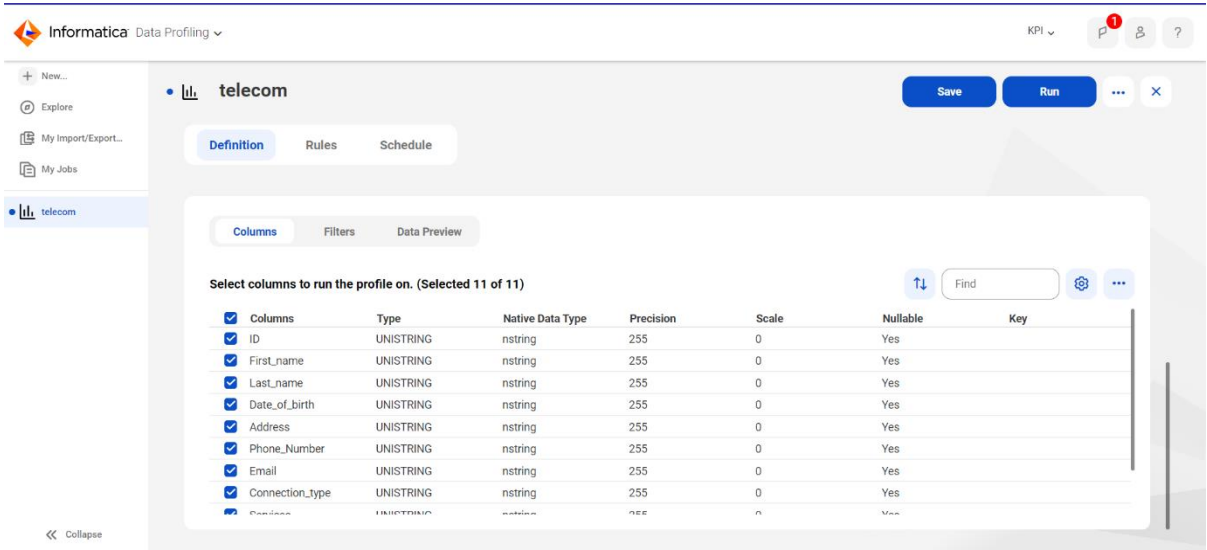


Рис.3.69 Прев’ю даних, які отримані з файла

Після проведення всіх налаштувань потрібно зберегти об’єкт і натиснути кнопку “Run” для того, щоб запустити його обробку. Після того як об’єкт відправлено на обробку, прогрес із його обробки відображається на вкладці “My Jobs” (рис.3.70).

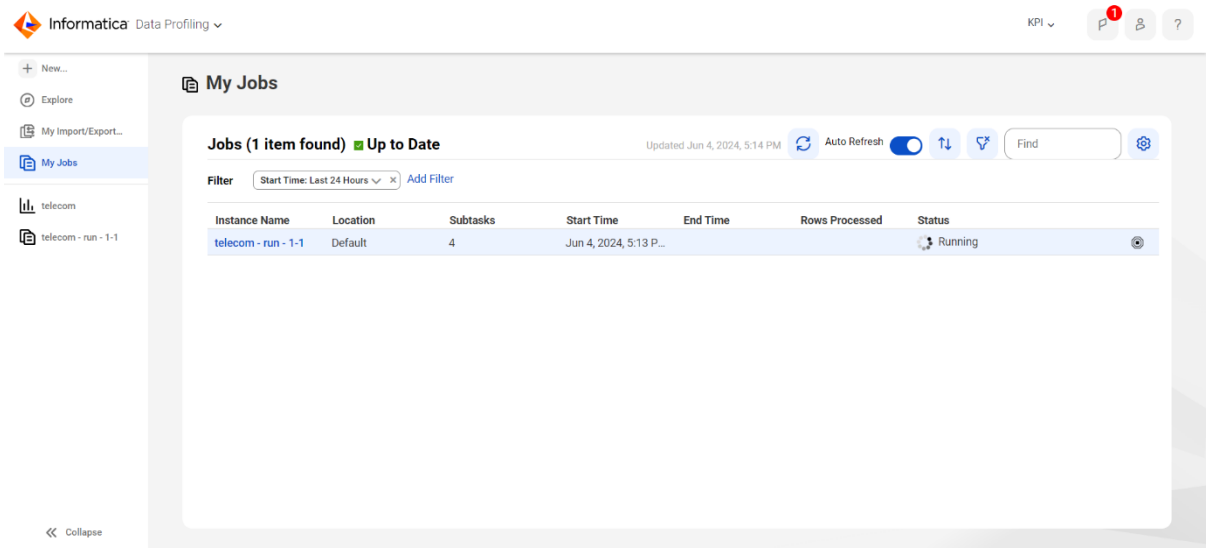


Рис.3.70 Вкладка із запущеними на виконання завданнями

Тут можна побачити, які завдання знаходяться на виконанні, коли вони були початі і закінчені, скільки є підзавдань, а також статус виконання (рис.3.71).

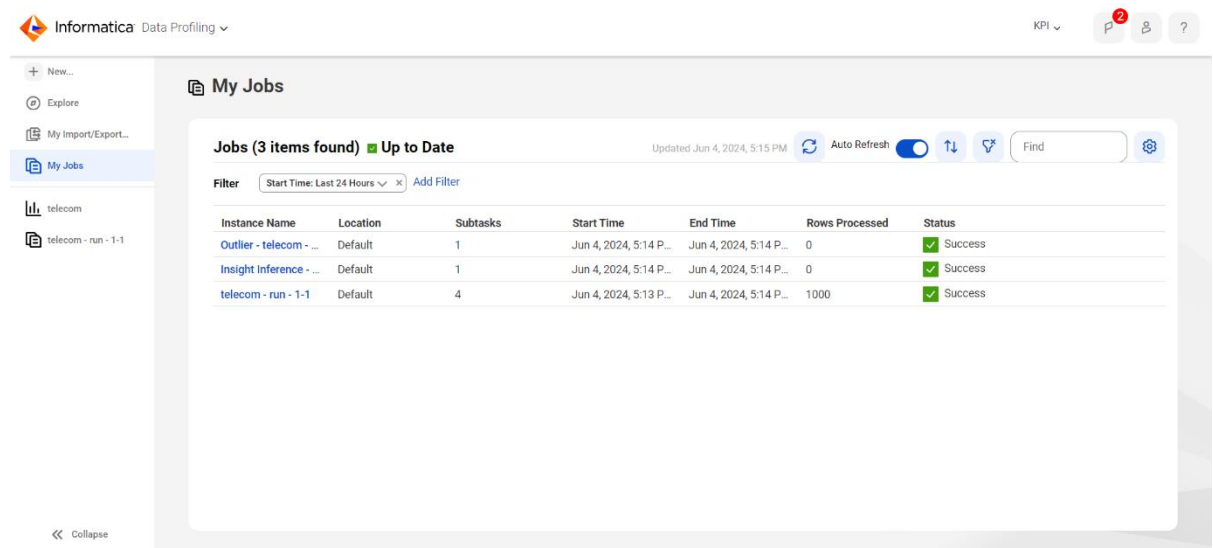


Рис.3.71 Успішно виконані всі завдання для профілювання

Після того як всі завдання успішно виконані, можна перейти по об'єкту Profiling, який був створений, і після цього натиснути на сповіщенні зверху, для того щоб оновити дані (рис.3.72).

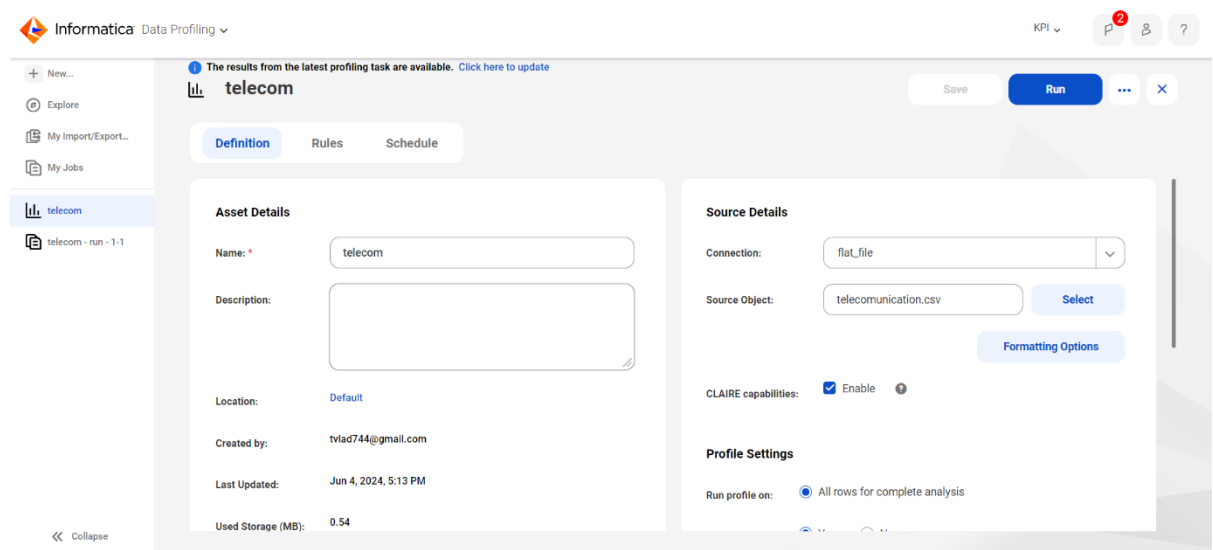


Рис.3.72 Сповіщення про доступність даних із профілювання

Після оновлення користувач потрапляє на сторінку з профільованими даними (рис.3.73). Тут як і в інших протестованих інструментах відображаються різні корисні дані для того, щоб розуміти загальну картину про якість цих даних. Також є функція аналізу патернів значень для полів. На рисунку нижче показано результат профілювання даних із датасету з валідними даними. Можна побачити, що хибних розпізнавань значень null немає, а тому можна говорити про те, що дані коректно завантажуються в інструмент.

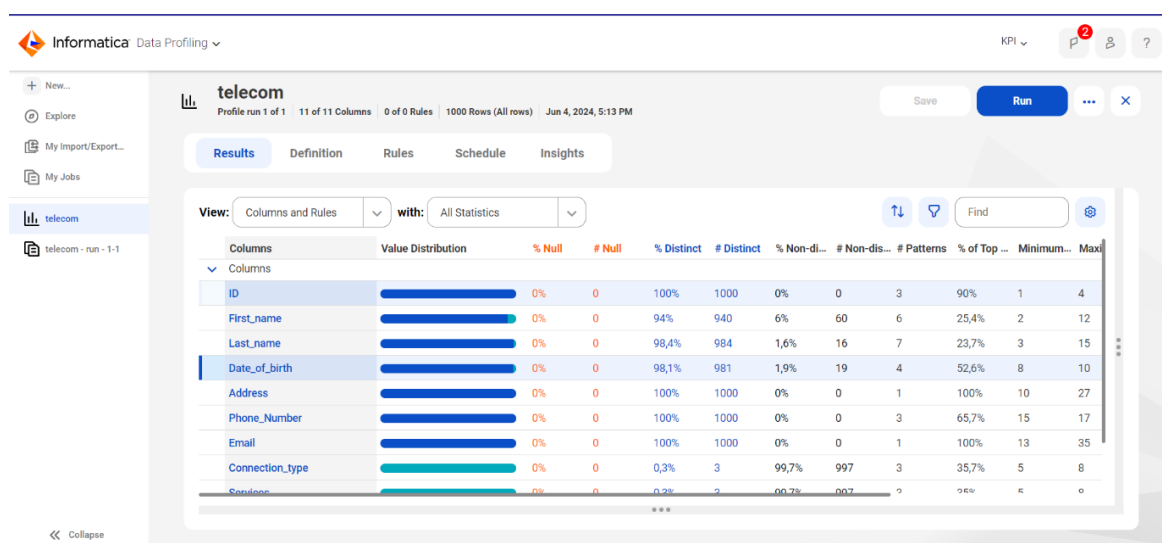


Рис.3.73 Результати профілювання

Також для порівняння були завантажені дані з датасету, який містить невалідні дані. Можна побачити, як відображається кількість значень Null на смужці розподілу значень (рис.3.74).

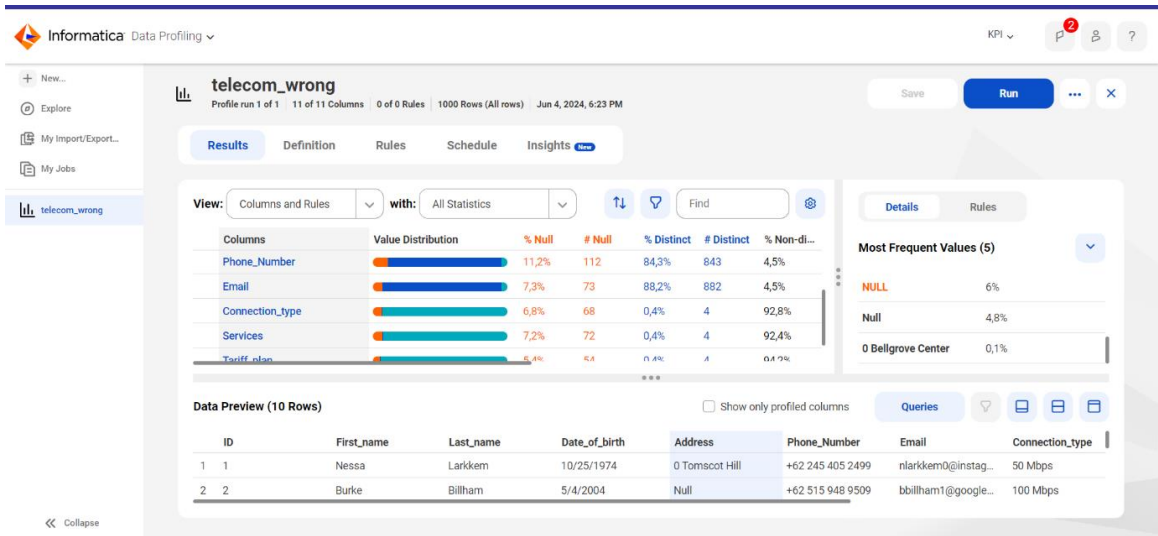


Рис.3.74 Результат профілювання датасету із пустими полями

Також при натисканні на ім'я конкретного поля користувач бачить детальнішу статистику по цьому полі й може побачити частоту певних значень на візуалізації, частоту значень по певному шаблону і ще багато статистичних даних. Також при профілюванні можна застосовувати певні правила специфікацій до конкретного поля, після чого буде відображено яка кількість даних підпала під ці специфікації, й буде проведено таке ж профілювання як для інших полів (рис.3.75).

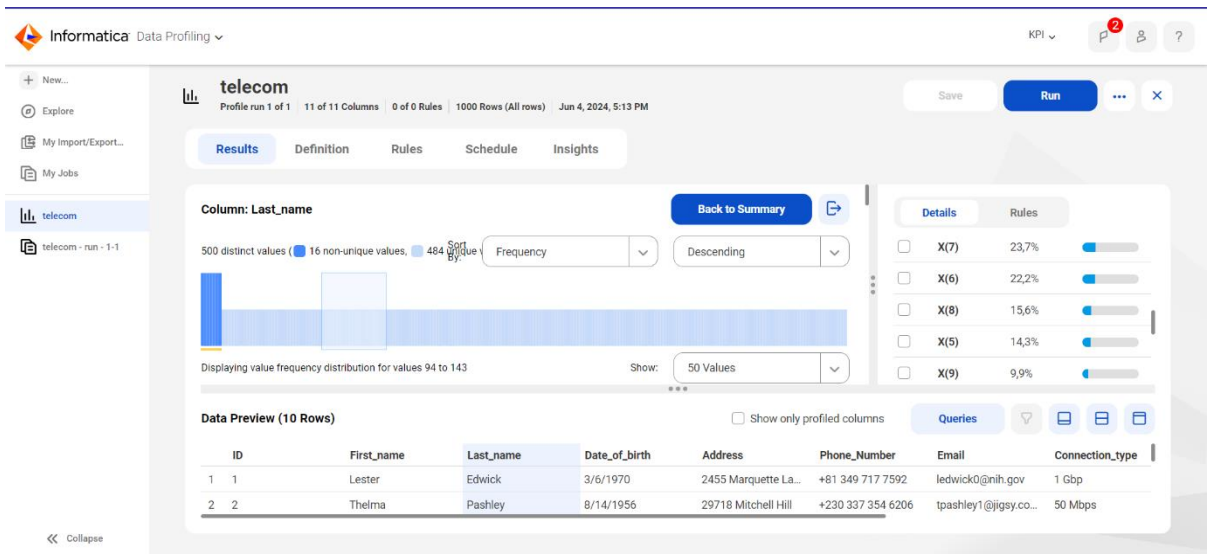


Рис.3.75 Детальні дані профілювання конкретного поля

Наступним важливим сервісом являється сервіс Data Quality (рис.3.76). Саме він являється основним предметом аналізу.

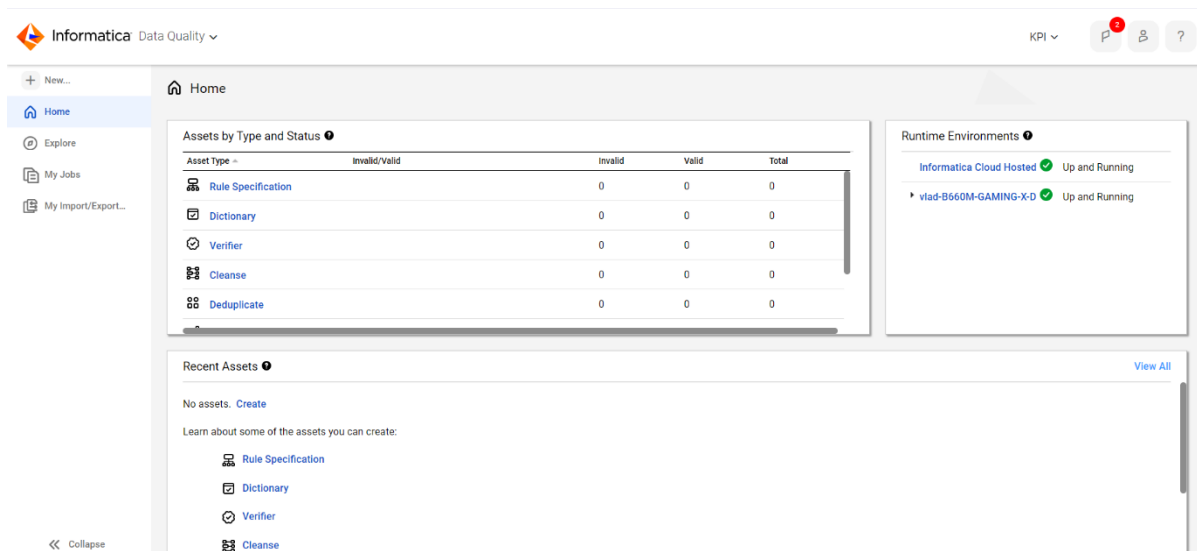


Рис.3.76 Головне вікно сервісу Data quality

В даному сервісі можна створити декілька об'єктів, які безпосередньо належать до інструментів валідації даних (рис.3.77).

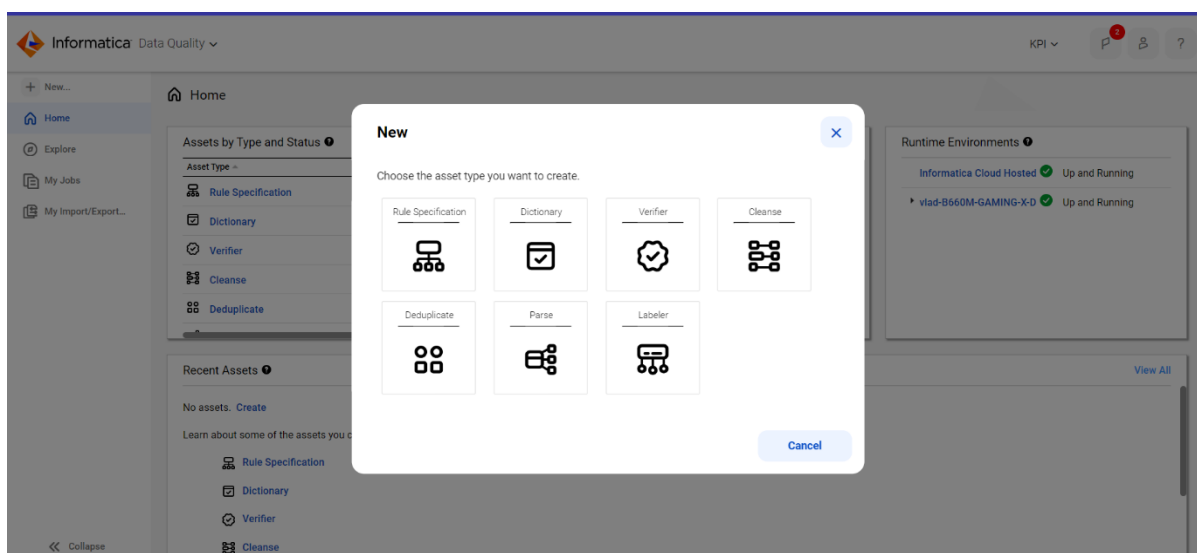


Рис.3.77 Об'єкти, які створюються в даному сервісі

Першим об’єктом є правила специфікацій. При створенні даного об’єкту потрібно назвати його, а також можна додати детальніший опис. Також є поле “Dimension”, в якому можна обрати, до перевірки якої метрики відноситься дане правило (рис.3.78).

Рис.3.78 Створення правила специфікації

У вкладці конфігурації даного об’єкта знаходиться інтерфейс створення конкретних специфікацій (рис.). Вони представлені у вигляді умов if then, або if then else. Є великий вибір у виборі операторів, а також умов, для перевірки яких також є великий перелік вбудованих функцій, таких як перевірка типів даних, отримання дня тижня із дати і купа інших. Також є великий перелік можливих дій як результат умови (рис.3.79).

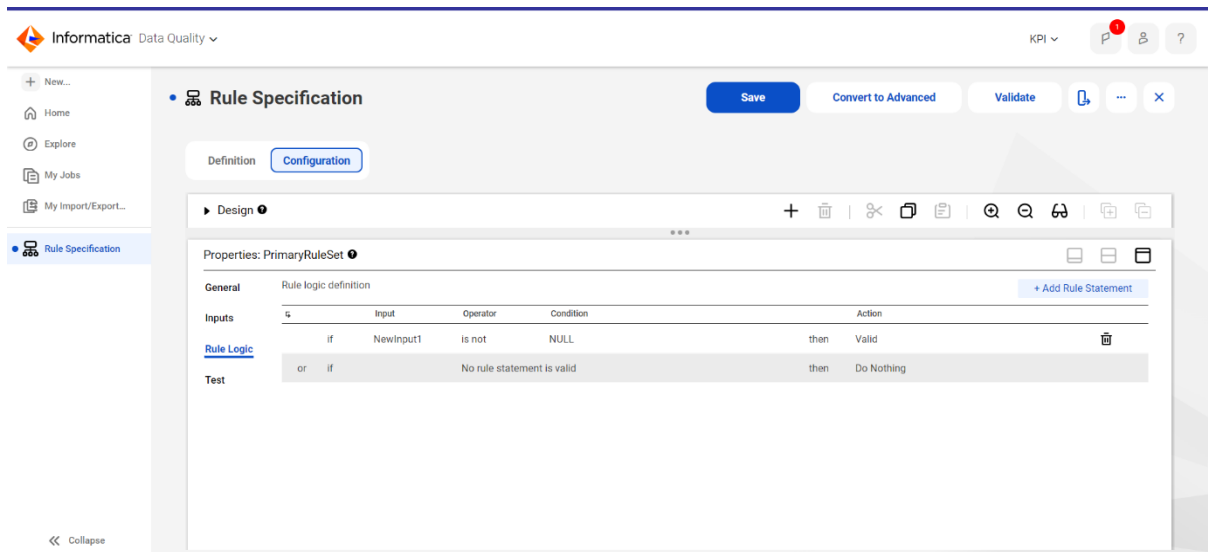


Рис.3.79 Створення логіки правила

Далі показано правило специфікацій, яке видаляє множинні пробіли в значенні поля (рис.3.80).

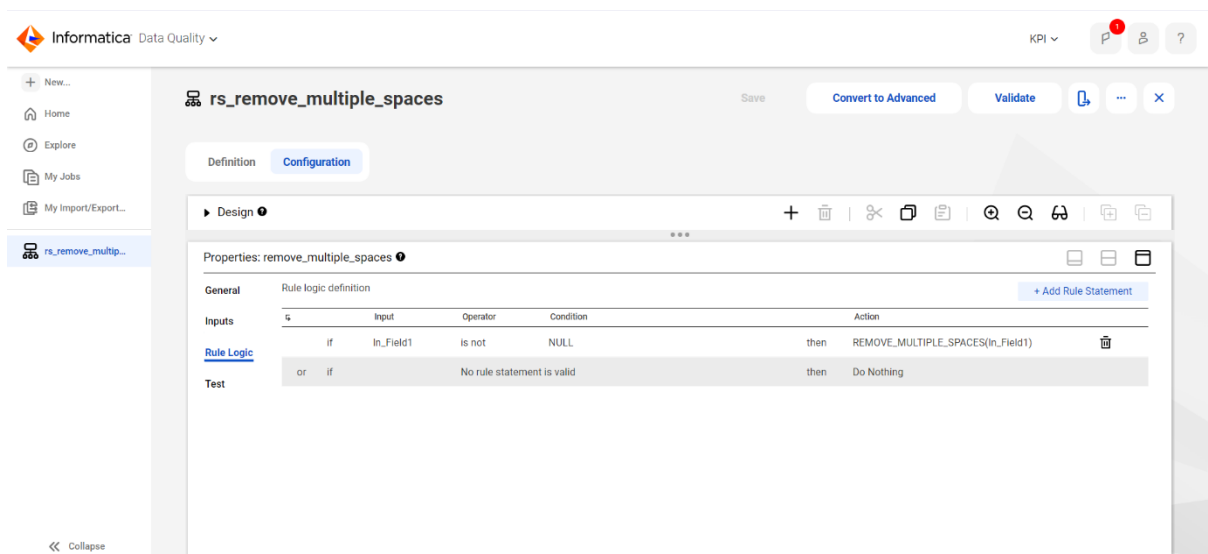


Рис.3.80 Приклад правила, що видаляє множинні відступи

Наступним об'єктом в даному сервісі є словники. Словники призначені для перевірки словникових значень, а також потенційної заміни збігів на значення з словника. На вкладці визначення також можна дати ім'я об'єкту та описати його (рис.3.81).

Informatica Data Quality

KPI

New... Home Explore My Jobs My Import/Export... Dictionary

Dictionary

Definition Configuration

Save

Asset Details

Name: * Dictionary

Description: Enter description...

Location: * DQcore Browse

Collapse

Рис.3.81 Створення словника

На вкладці конфігурації потрібно вводити значення, які будуть в словнику, або можна імпортувати дані в даний словник із зовнішніх налаштованих підключень (рис.3.82).

Informatica Data Quality

KPI

New... Home Explore My Jobs My Import/Export... Dictionary

Dictionary

Definition Configuration

Save

Items (50)

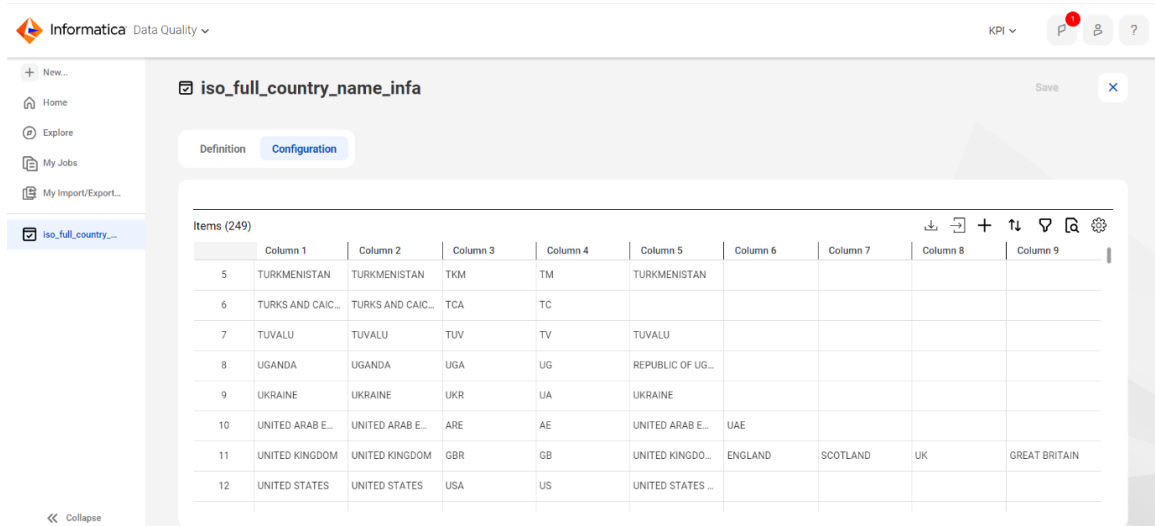
	Column 1	Column 2	Column 3	Column 4	Column 5
1					
2					
3					
4					
5					
6					
7					
8					

Collapse

Рис.3.82 Налаштування словника

На рисунку нижче показано приклад словника, який містить скорочення країн, а також їхні повні назви. Цей словник може використовуватись для

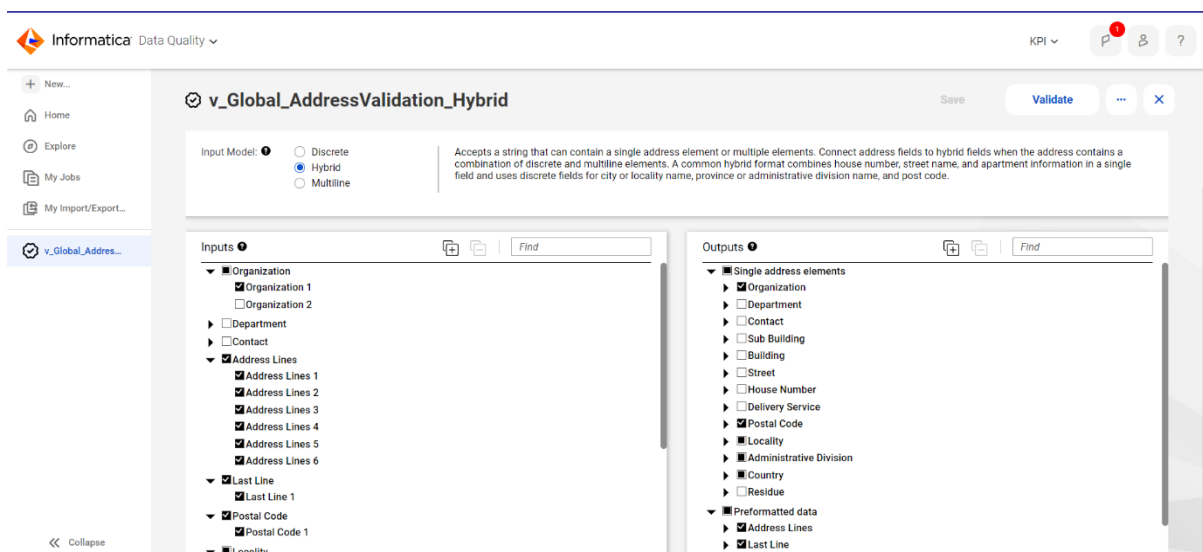
стандартизації значень, тобто приведення значень в цьому полі до єдиного стандартизованого вигляду (рис.3.83).



	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9
5	TURKMENISTAN	TURKMENISTAN	TKM	TM	TURKMENISTAN				
6	TURKS AND CAIC...	TURKS AND CAIC...	TCA	TC					
7	TUVALU	TUVALU	TUV	TV	TUVALU				
8	UGANDA	UGANDA	UGA	UG	REPUBLIC OF UG...				
9	UKRAINE	UKRAINE	UKR	UA	UKRAINE				
10	UNITED ARAB E...	UNITED ARAB E...	ARE	AE	UNITED ARAB E...	UAE			
11	UNITED KINGDOM	UNITED KINGDOM	GBR	GB	UNITED KINGDO...	ENGLAND	SCOTLAND	UK	GREAT BRITAIN
12	UNITED STATES	UNITED STATES	USA	US	UNITED STATES ...				

Рис.3.83 Приклад словника із скороченнями назв країн та їх повними назвами

Наступним об'єктом є верифікатор. Даний об'єкт потрібен для верифікації поля фізичної адреси на точність та доступність. Також даний об'єкт може виправляти та в деяких випадках доповнювати адресу. Нижче показано приклад налаштувань верифікатора для перевірки гібридного поля адреси (рис.3.84).



Input Model: ☐ Discrete ☒ Hybrid ☐ Multiline

Accepts a string that can contain a single address element or multiple elements. Connect address fields to hybrid fields when the address contains a combination of discrete and multiline elements. A common hybrid format combines house number, street name, and apartment information in a single field and uses discrete fields for city or locality name, province or administrative division name, and post code.

Inputs:

- Organization
 - ☒ Organization 1
 - ☐ Organization 2
- Department
- Contact
- Address Lines
 - ☒ Address Lines 1
 - ☒ Address Lines 2
 - ☒ Address Lines 3
 - ☒ Address Lines 4
 - ☒ Address Lines 5
 - ☒ Address Lines 6
- Last Line
 - ☒ Last Line 1
- Postal Code
 - ☒ Postal Code 1
- Locality

Outputs:

- Single address elements
 - ☒ Organization
 - ☐ Department
 - ☐ Contact
 - ☐ Sub Building
 - ☐ Building
 - ☐ Street
 - ☐ House Number
 - ☐ Delivery Service
 - ☒ Postal Code
 - ☒ Locality
 - ☒ Administrative Division
 - ☐ Country
 - ☐ Residue
- Preformatted data
 - ☒ Address Lines
 - ☒ Last Line

Рис.3.84 Приклад верифікатора для перевірки можливості доставки до певної адреси з різних країн

Наступним йде об'єкт очищення. За допомогою нього можна міняти регістр текстових полів, видаляти певні значення та зайві пропуски, а також можна замінювати одні значення на інші(рис.3.85). Саме в цьому об'єкті можна використовувати раніше створені словники для стандартизації даних

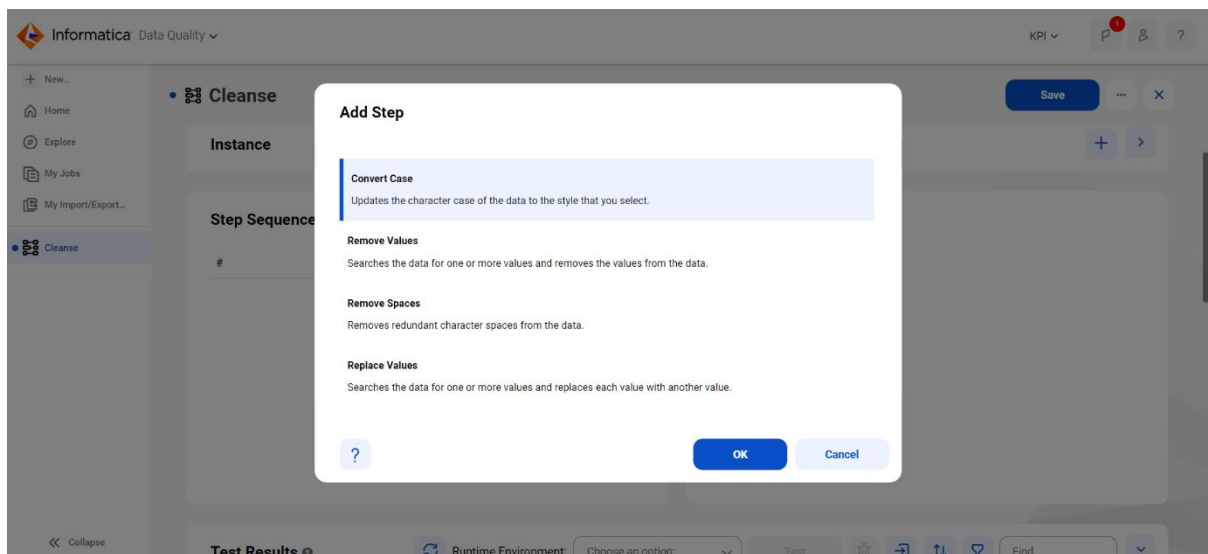


Рис.3.85 Можливі дії в очиснику

Далі показано приклад очисника, який міняє регістр тексту в полях. При цьому можна вказати, який роздільник використовується в даних (рис.3.86).

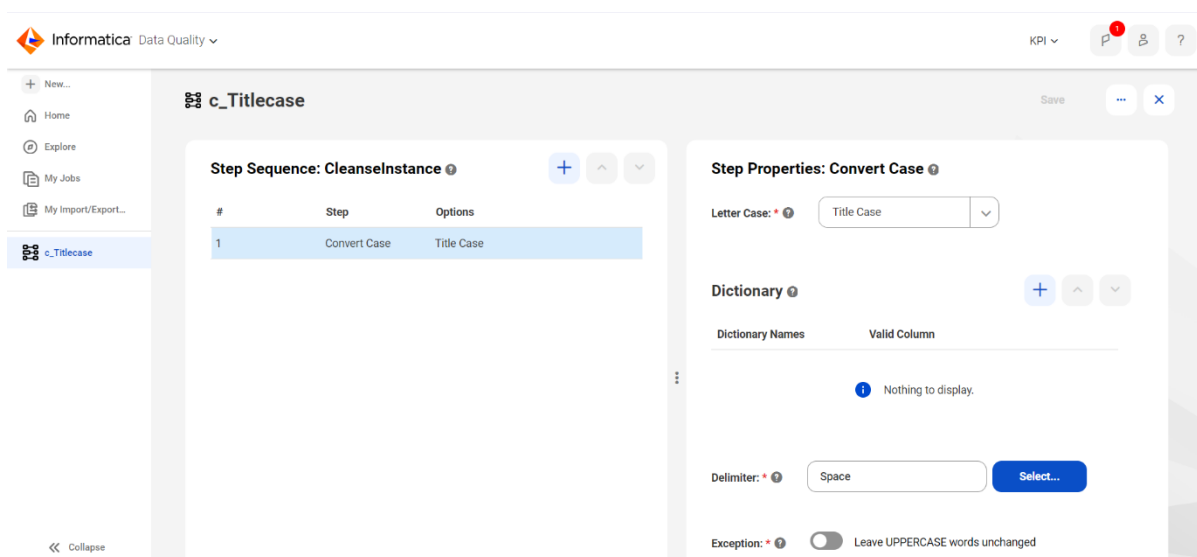


Рис.3.86 Приклад очисника, який змінює регістр на заголовочний

Також майже кожен об'єкт із даного сервісу можна відразу після налаштування тестувати, для того щоб оцінити поведінку створеного об'єкта і мати змогу швидко редагувати створюваний об'єкт (рис.3.87).

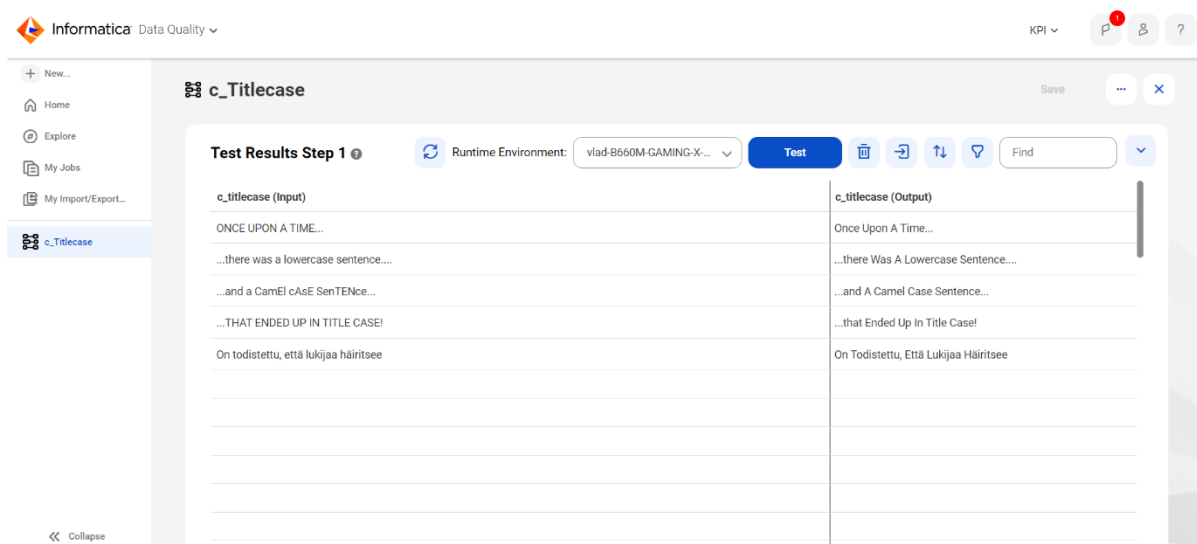


Рис.3.87 Тестування очисника

Наступним об'єктом є дедублікатор. Даний об'єкт призначений для налаштування пошуку та поведінки з дублікатами. При налаштуванні потрібно обрати тип інформації, в якій буде проводитись пошук дублікатів, ключ

індексації, який індексує дані в цьому полі, регіон, з якого зібрані дані, а також режим продуктивності, який залежить від порогу часу, який припустимо витратити на дану операцію (рис.3.88).

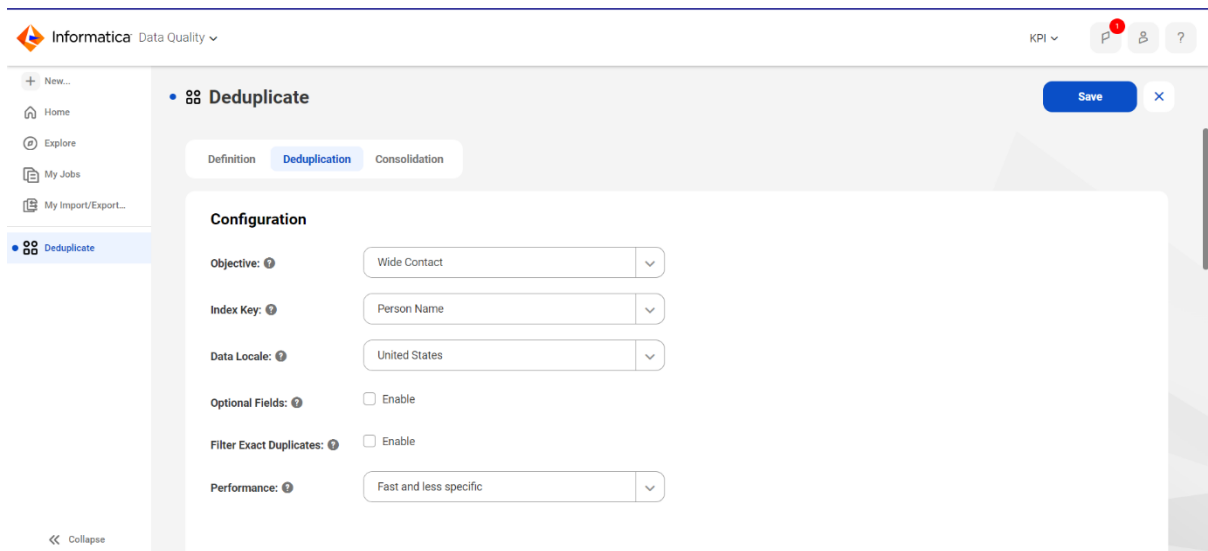


Рис.3.88 Налаштування дедублікатора

Також додатково можна налаштувати допустимий рівень схожості, рівень точності, рівень впевненості, а також формат виводу дублікатів (рис.3.89)

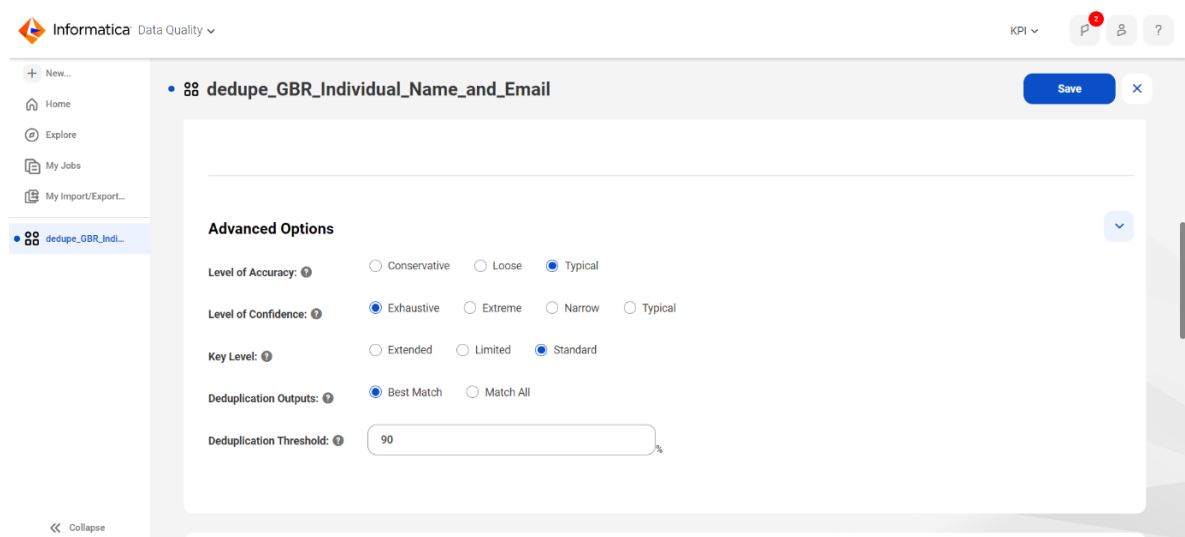


Рис.3.89 Розширені налаштування дедублікатора

Наступним об'єктом є парсер. Він потрібен для того щоб аналізувати та розбивати значення полів на декілька значень. Це може застосовуватись коли є комплексні поля, такі як ПІБ, повна адреса, електронна адреса, і для зручності аналізу можна парсити ці дані на окремі значення, після чого аналізувати їх (рис.3.90).

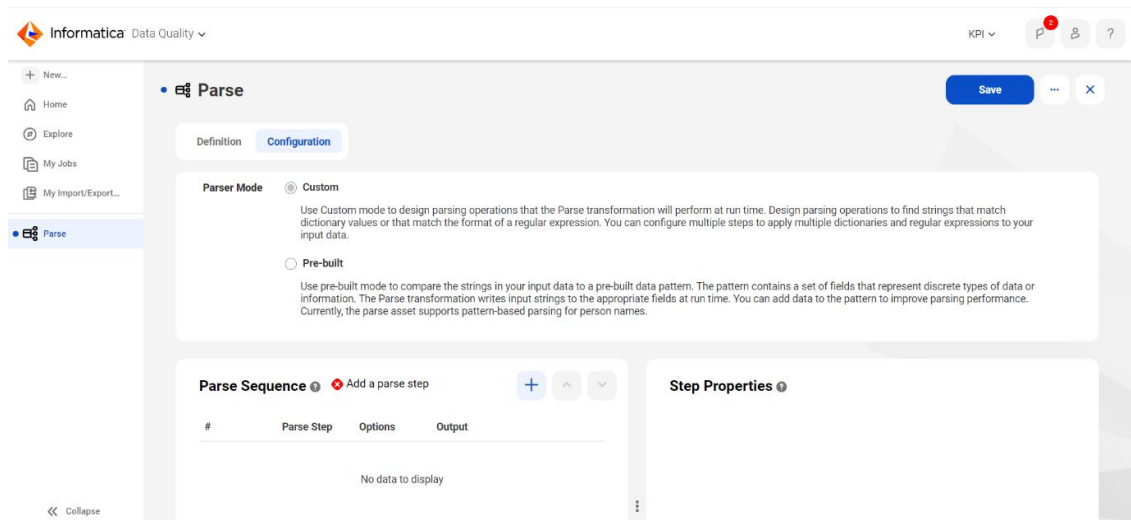


Рис.3.91 Налаштування парсера

Далі показано приклад парсера, який відділяє від адреси електронної пошти частину до знаку “@” (рис.3.92).

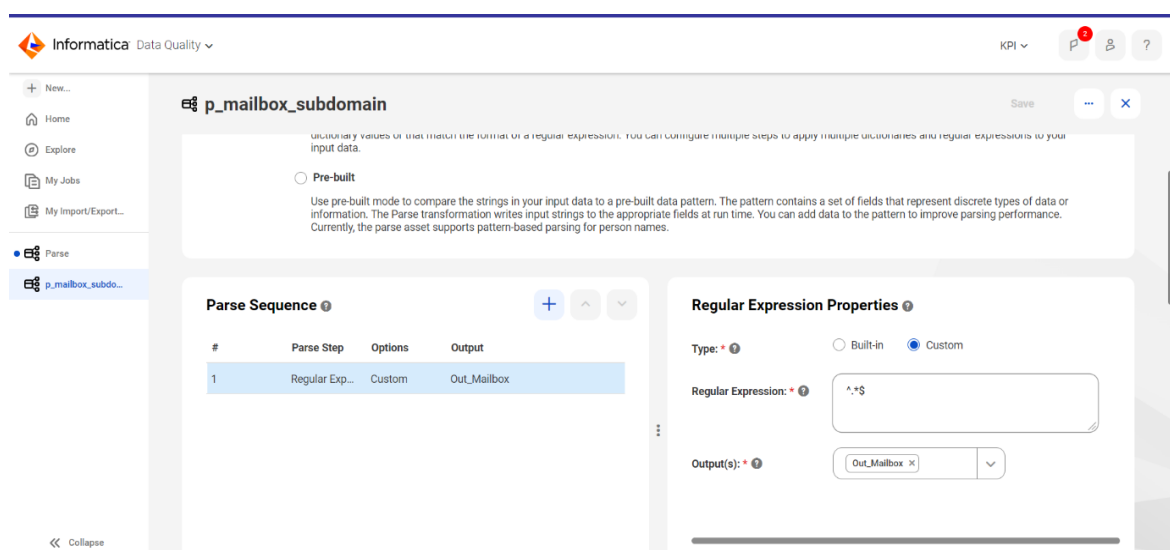


Рис.3.92 Приклад парсера субдомена електронної пошти

Наступним і останнім об'єктом є маркувальник. Він маркує вказані користувачем символи в полях вказаним символом. Тобто якщо розглядати цей об'єкт простіше, то він робить заміну символів, але так як цей об'єкт призначений саме для цього, його швидше та зручніше налаштовувати, хоча цю дію можна виконувати й через інші об'єкти (рис.3.93).

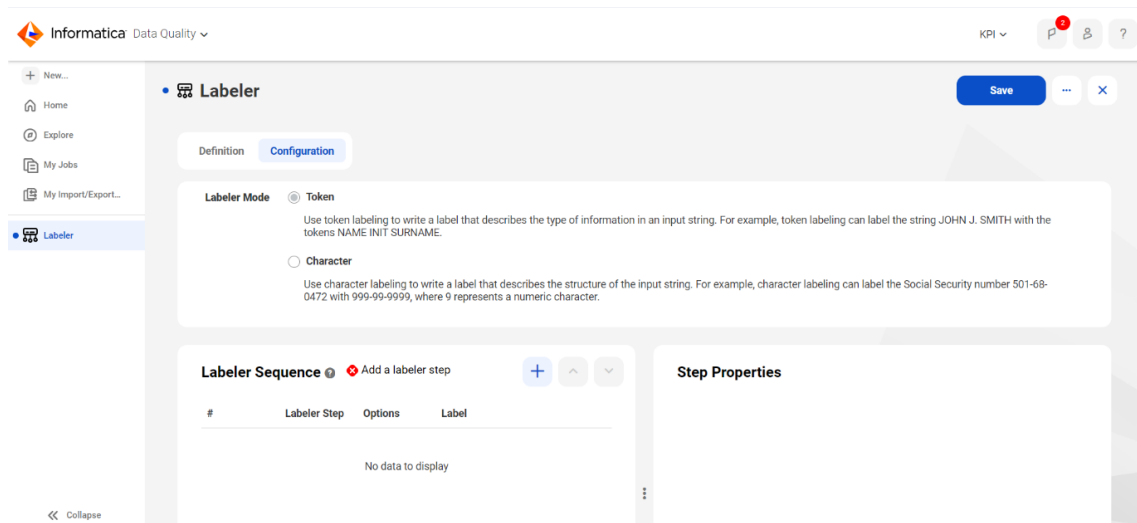


Рис.3.93 Налаштування маркувальника

Можна також побачити налаштований маркувальний, який замінює в полях символи з переліку, на літеру “S” (рис.3.94).

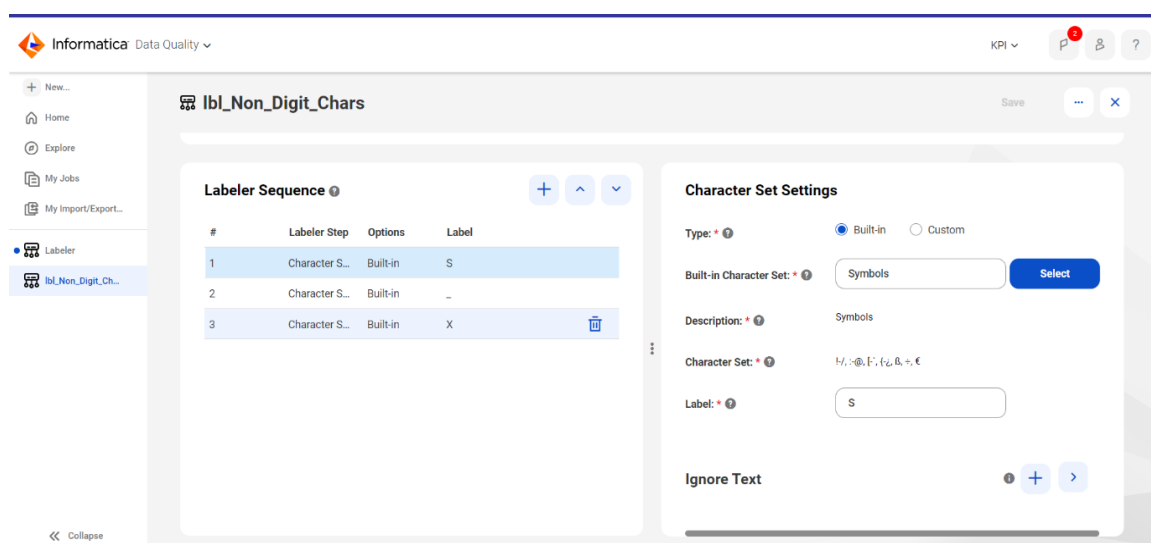


Рис.3.94 Приклад налаштованого маркувальника

Наступним важливим сервісом є Data Integration. Цей сервіс призначений більше для інтеграції та обробки даних, завдяки об'єктам які були створені в сервісі Data Quality і не тільки. На головному вікні відображаються зведені дані щодо стану асетів, середовищ виконання, стан запущених завдань а також нещодавні об'єкти (рис.3.95).

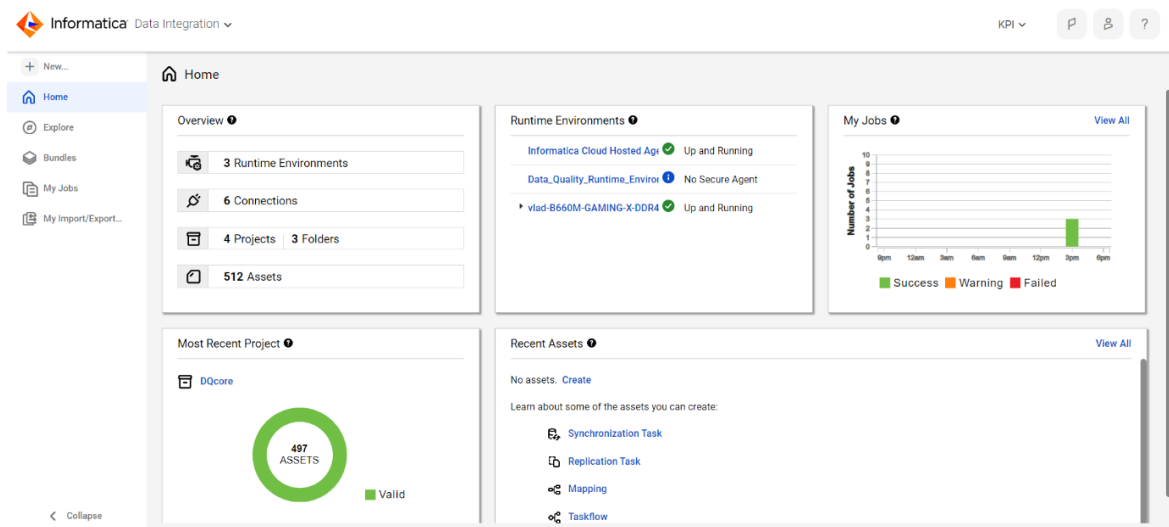


Рис.3.95 Головний екран сервісу інтеграції даних

В даному сервісі можна створити безліч об'єктів, таких як мапінги, маплети (користувачські блоки, які створюються зі звичайних блоків з метою спрощення використання певних шаблонів блоків), робочі процеси, певні нестандартні компоненти, а також завдання, які створюються з раніше створених мапінгів (рис.3.96).

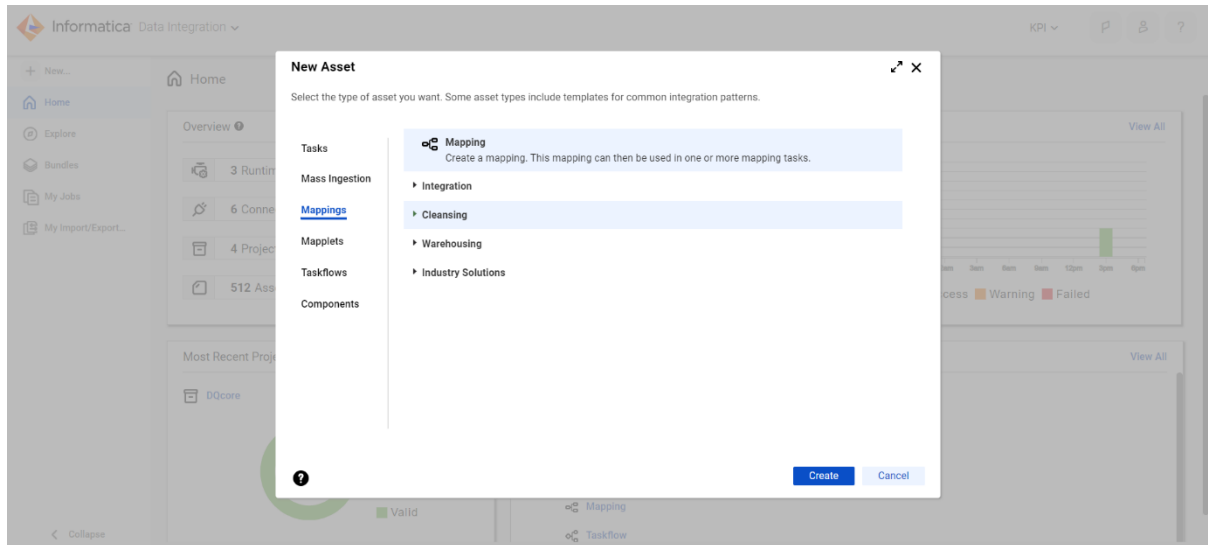


Рис.3.96 Об'єкти, які можна створити в даному сервісі

На рисунку нижче можна побачити, як в цілому виглядає створення мапінга. Створення відбувається із стандартних блоків, кожен з яких по своєму налаштовується, і після цього ці блоки з'єднуються в потрібному порядку для виконання раніше визначених функцій. Як приклад, за допомогою мапінгів можна завантажити дані з джерела, перевірити ці дані по певному правилу специфікацій створеному в сервісі Data Quality, і на виході з перевіркою по правилу дані відправляються різними гілками, в яким дані знову ж таки обробляються потрібним способом. На виході такі дані можна зберегти в будь яке джерело, з яким налаштоване з'єднання (рис.3.97).

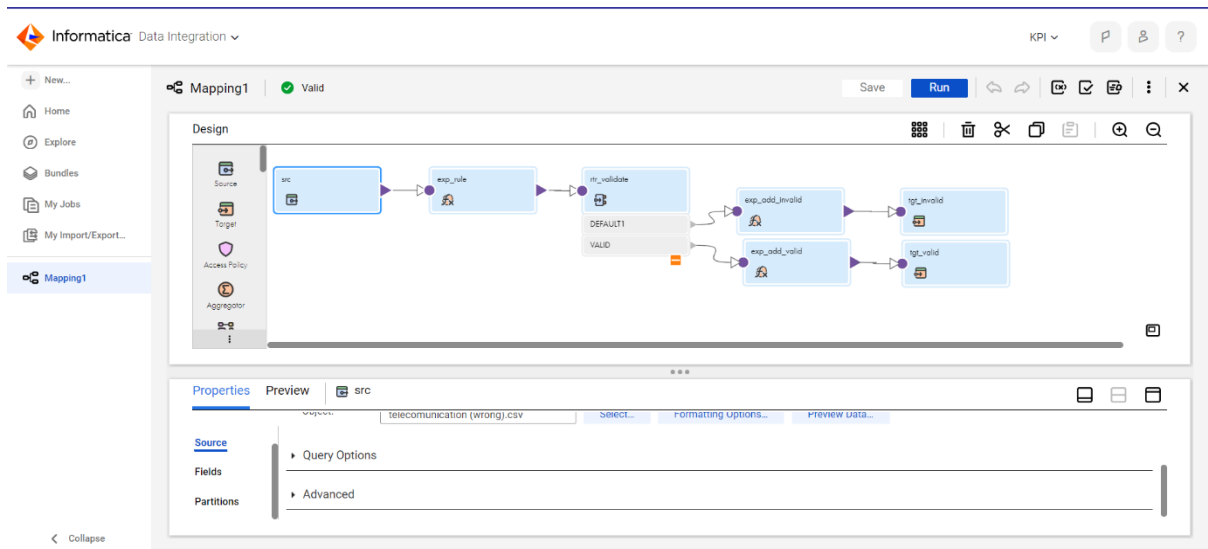


Рис.3.97 Приклад блок схеми для відокремлення даних з пустими полями та null

Проведемо тестування створеного датасету із згенерованими даними та штучно введеними помилками за допомогою даного інструменту. Фінальний вигляд схеми робочого процесу для обробки всіх помилок показано нижче (рис.3.98)

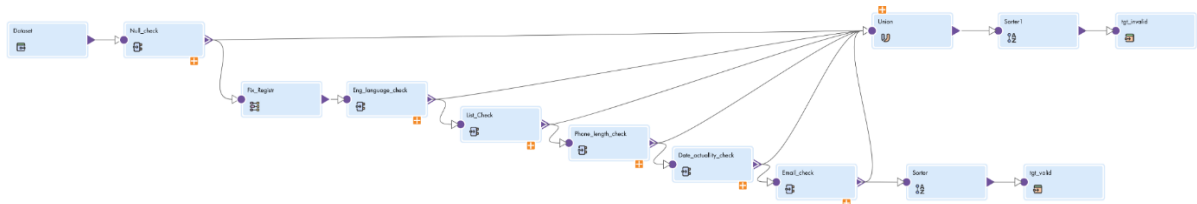


Рис.3.98 Схема робочого процесу для обробки помилок

Розберемо детальніше схему. На початку схеми присутній блок Source, завдяки якому отримуються дані з джерела. Наступним йде фільтр Null_check, який за допомогою умови, представленій на рисунку , відокремлює значення Null (рис.3.99).

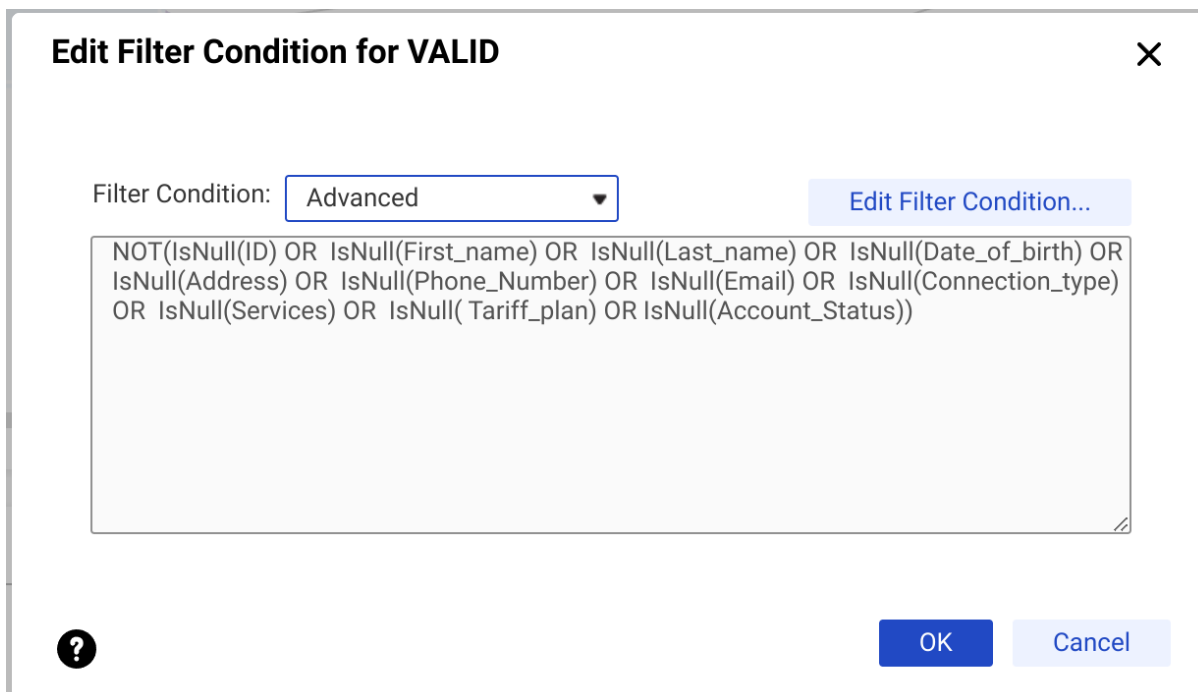


Рис.3.99 Фільтр Null_check

Наступним блоком є блок очищення Fix_registr, який може очищати значення. В моєму випадку це зручно для виправлення точності даних шляхом зміни регістру значень на вказаний. Налаштований блок очищення для зміни регістру на Рядковий показано нижче (рис.3.100).

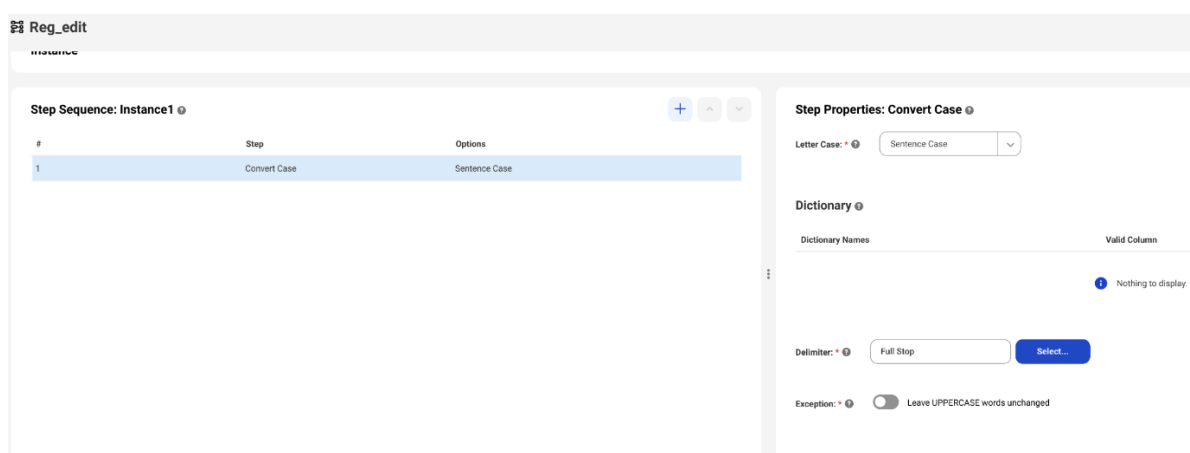


Рис.3.100 Блок очищення для зміни регістру на Рядковий

Наступним йде фільтр Eng_language_check, в якому за допомогою regex патерна відокремлюються значення полів First name, Last name та Address, які

містять символи, які не є не англійськими символами, тобто якщо простіше, то ми детектимо, чи складається значення лише з англійських букв та символів (рис.3.101).

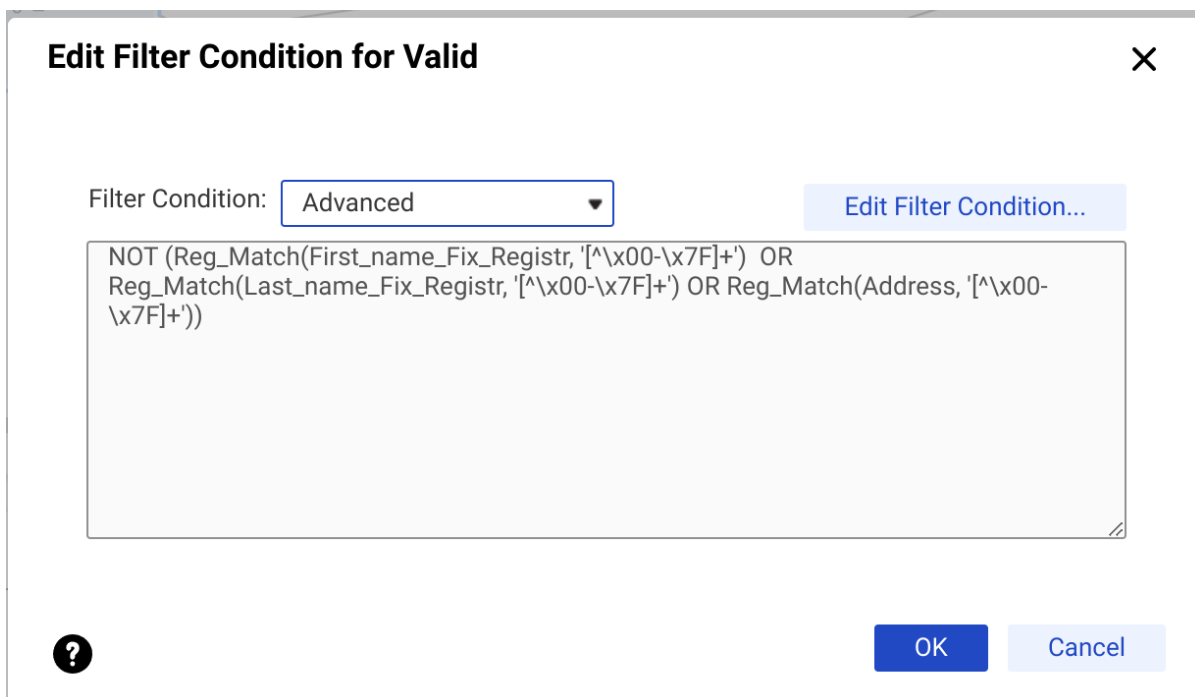


Рис.3.101 Фільтр Eng_language_check

Наступним є також фільтр Lisy_check, в якому жорстко заковано, які значення в яких полях повинні бути, і на виході ми отримуємо рядки, в яких значення полів Services, Connection_type, Tariff_plan та Account_status мають значення, перелічені в умові (рис.3.102).

Edit Filter Condition for Valid [X]

Filter Condition: Advanced ▼ Edit Filter Condition...

(Services = 'Cable TV' OR Services = 'Phone' OR Services = 'Internet') AND
 (Connection_type = '50 Mbps' OR Connection_type = '1 Gbp' OR Connection_type = '100 Mbps') AND (Tariff_plan = 'Premium Package' OR Tariff_plan = 'Basic Plan' OR Tariff_plan = 'Gold Plan') AND (Account_Status = 'Disconnected' OR Account_Status = 'Blocked' OR Account_Status = 'Active')

[?] OK Cancel

Рис.3.102 Фільтр Lisy_check

Наступним йде фільтр Phone_length_check, в якому перевіряється довжина поля Phone_number на відповідність умові, що довжина цього поля, враховуючи знак '+' та пробіли має бути не меншою за 15 та не більшою за 17 включно (рис.3.103).

Edit Filter Condition for Valid [X]

Filter Condition: Advanced ▼ Edit Filter Condition...

NOT(Length(Phone_Number) < 15 OR Length(Phone_Number) > 17)

[?] OK Cancel

Рис.3.103 Фільтр Phone_lenght_check

Наступним блоком є фільтр `Date_actually_check`, в якому перевіряється різниця між роком дати народження користувача та теперішньою датою (рис.3.104). За умовою актуальності, поставленою мною, задля забезпечення актуальності даних дана різниця не має бути більшою ніж 110 років, а також задля забезпечення достовірності, дана різниця не може бути меншою за 5 років.

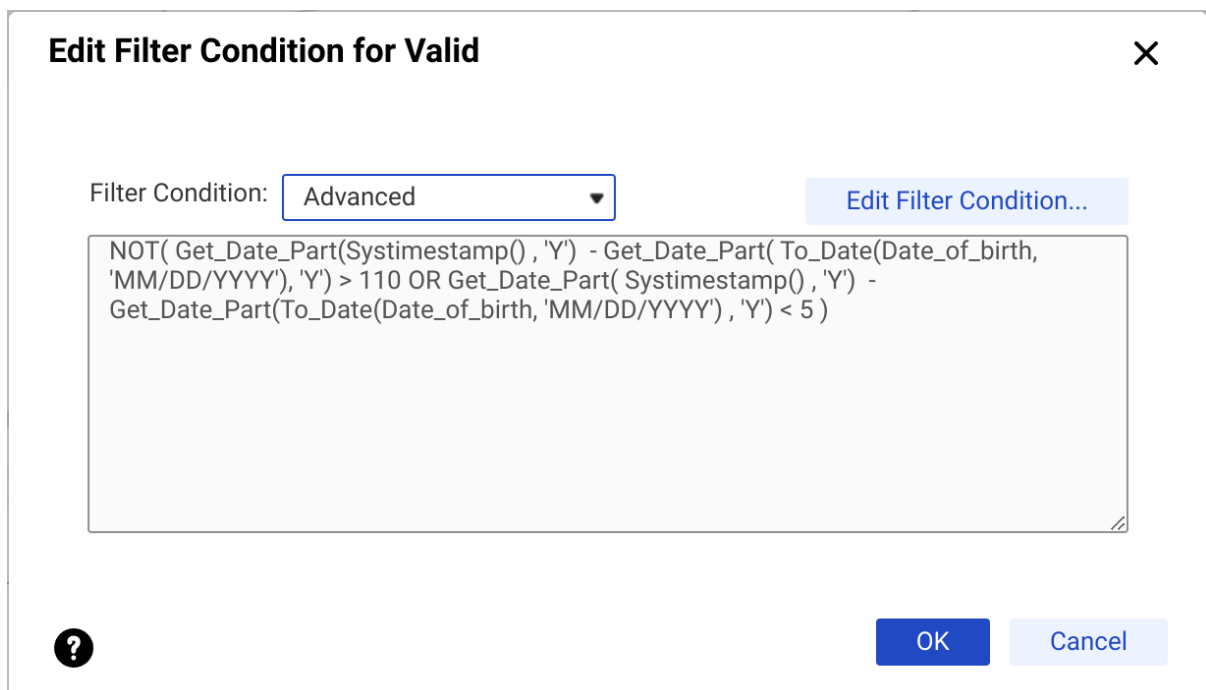


Рис.3.104 Фільтр `Date_actually_check`

Наступним йде фільтр `Email_check`, в якому за допомогою патерну `regex` перевіряється відповідність значень поля `Email` патерну електронних пошт, використаних в моєму датасеті (рис.3.105).

?

Edit Filter Condition for NewGroup

×

Filter Condition:

Advanced

Edit Filter Condition...

Reg_Match(Email, '.*?@.*?\..*')

OK

Cancel

Filter Condition: **Advanced**

Edit Filter Condition...

```
Reg_Match(Email, '.*?@.*?\..*')
```

OK

Cancel

Рис.3.105 Фільтр Email_check

Після обробки всіма цими блоками я вивела валідні та невалідні дані в окремі файли. Невалідні дані зображені нижче (рис.3.106).

ID	First_name	Last_name	Date_of_birth	Address	Phone_Number	Email	Connection_type	Services	Tariff_plan	Account_Status
9	Marcia	Downgate	1/17/1942	9768 Thackeray Center	+63 201 430 6893	mdowngate8@toplist.cz	1 Gbp	Smartphone	Basic Plan	Blocked
18	Husein	Pellet	2/16/1969	386 Dryden Trail	+33 520 525 5592	hpellet@marketwatch.com	1 Gbp	Internet	Basic Plan	
39	Celene	Chismon	7/15/1962	77 Doe Crossing Place	+86 625 426 3921	ccwismon12@macromedia.com	1 Gbp	Cable TV	Premium Package	Активный
54	Cecelia	Trewett	6/4/1993	7640 Cambridge Point	+502 979 416 2212	ctrewett1@mac.com	100 Mbps		Basic Plan	Disconnected
68	Lindsay	Hedin	07/05/2050	74 Boyd Avenue	+62 732 280 3655	lhedin1v@don.ne.jp	100 Mbps	Internet	Premium Package	Active
85	Earvin	Majury	9/16/1992	94 Swallow Street	+7 500 212 4355	emajury2c@squidoo.com	100 Mbps	Internet	Gold Plan	Active
161	Bastien	Pauleit	01/06/1900	3 Toban Parkway	+86 641 181 0311	bpauleit4g@skype.com	50 Mbps	Phone	Basic Plan	Disconnected
164	Candide	Marschke	9/16/2014	265 TERNYSON ALLEY	+62 291 350 7319	cmarschke41@amazon.co.jp	50 Mbps	Phone	Premium Package	Блокировка
293	Cointon	Bar	7/29/1972	7 Hanson Street	+93 406 959 8434	cber84@goo.gl	50 Mbps	Internet	Minimum Plan	Blocked
343	Kora	Merioth	6/13/1952	74955 Magdeline Circle	+502 914	kmerioth9l@imgur.com	50 Mbps	Cable TV	Basic Plan	Disconnected
438	Courtney	Anyene	9/27/1994	3 Rusk Terrace	+375 684 228 1421 000	carylene5c@redcross.org	50 Mbps	Cable TV	Premium Package	Disconnected
509	Sibyl	Pashby	10/29/2007	9975 Mallard Street	+380 221 323 6833	spashbye4@reverbnation.com	one Gbp	Phone	Basic Plan	Blocked
641	Willie	Braghini	3/23/1988	79098 Hermina Point	+380 880 886 2796	wbraghinih@booking.com	one hundred Mbps	Internet	Gold Plan	Active

Рис.3.106 Вивід невалідних даних

Рядок з файлу із валідними даними, в якому можна побачити, що реєстр значення First name було успішно виправлено (рис.3.107).

7	Lacy	Hymus	1/30/2018	90272 Grasskamp Point	+86 930 740 0467	lhymus6@t.co	1 Gbp	Internet	Gold Plan	Blocked
---	------	-------	-----------	-----------------------	------------------	--------------	-------	----------	-----------	---------

Рис.3.107 Виправлення реєстру значення First_name

Результатом обробки мого датасету з штучно введеними помилками є обробка всіх 14 помилок, і одну з помилок вдалось легко виправити, через що виправлений рядок було включено до валідних даних. Тобто можна зробити висновки, що даний інструмент вдалось налаштувати на виявлення всіх штучно зроблених помилок, які порушували атрибути якості даних.

Даний інструмент має різку криву навчання. На початку використання з даним інструментом тяжко взаємодіяти через велику кількість функціоналу, а також незвичайний розподіл по сервісам й неінтуїтивність функціоналу певних об'єктів. Також спочатку тяжко розібратись із документацією, так як вона розкидана по великій кількості різних сторінок, файлів та відеогайдів. Але після того як приходить розуміння, що і як працює, інструментом приємно користуватись. Він має зручний розподіл, велику кількість заготовлених асетів, велику кількість курсів для навчання (Хоча велика кількість курсів є платною, також наявні хороші безкоштовні курси).

В цілому інструмент себе показав із хорошої сторони як інструмент для валідації даних, через великий функціонал, підтримку великої кількості джерел отримання даних, підтримку інтеграції з іншими інструментами та за швидкість виконання тестування або інших операцій. Також при спробі обробки власного датасету з штучно введеними помилками інструмент вдалось налаштувати так, щоб він зміг обробити всі 14 помилок, які порушували атрибути якості даних.

Окрім протестованих вище тулів, має місце перевірка якості даних за допомогою штучного інтелекту. Зараз такі системи стають дедалі популярнішими, і впроваджують дедалі більшою кількістю сервісів. Через таку популярність штучного інтелекту дедалі більше людей задумуються про те, що штучний інтелект зможе замінити їх на певній посаді. Тож варто перевірити, чи зможе сучасний штучний інтелект впоратись із перевіркою якості даних.

Перевірку я буду проводити за допомогою найпопулярнішого чат-боту із штучним інтелектом ChatGPT. Для тестування даних буде використовуватись найновіша версія даної моделі, а саме gpt-4o.

Для початку запитую, чи знає він, що таке стандарт якості даних ISO 25012. Як видно, дана модель знає що це за стандарт та які його основні положення (рис.3.108).

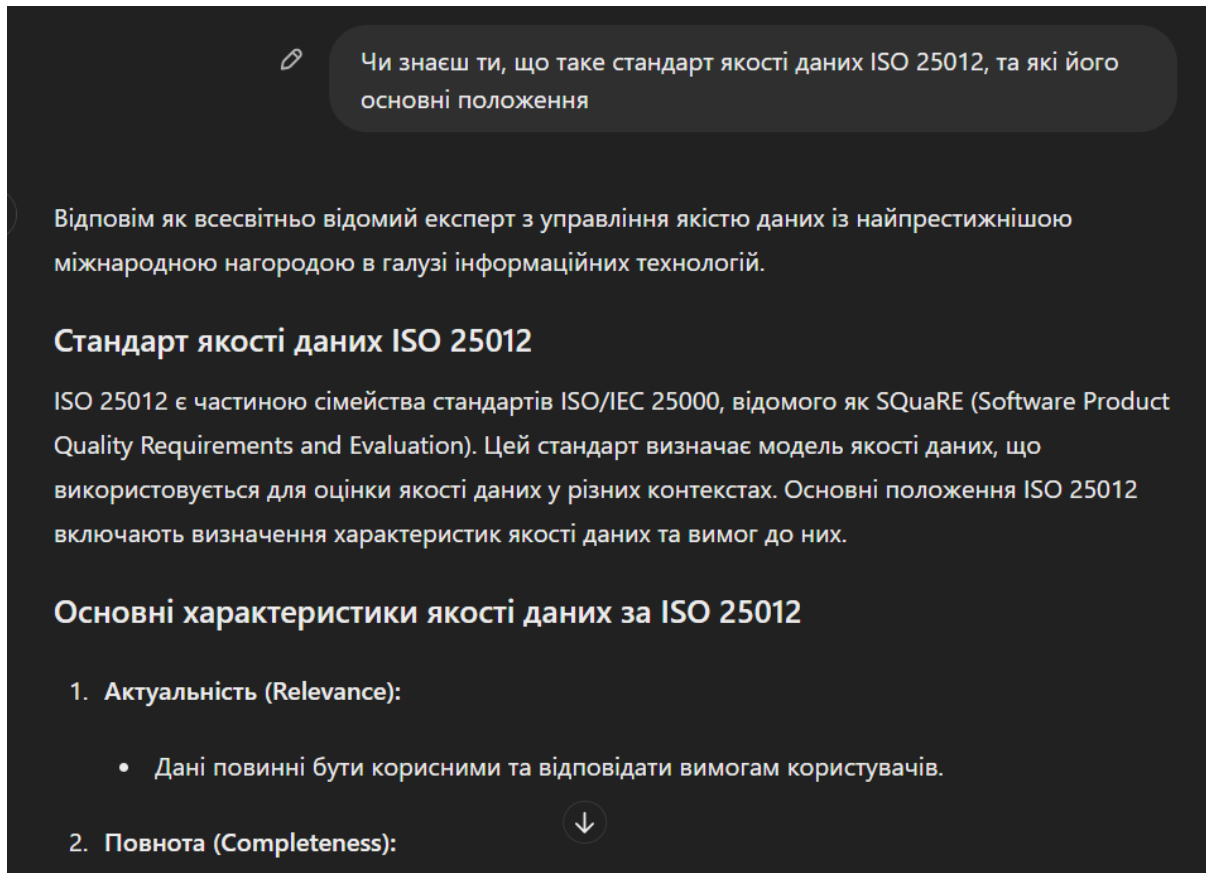


Рис.3.108 Знання чату gpt про стандарт ISO 25012

Наступним кроком є перевірка того, наскільки добре дана модель справиться з перевіркою даних на основі раніше створеного датасету. Починаю із простого запиту, без конкретизації певних аспектів (рис.3.109)

Telecommunication06			
	ID;First_name;Last_name		
1	1;Rita;Reye;12/19/1943;42 Roth Center;812 313 305 460;rreye0@zdneta.com;1		
2	2;Dag;Abelson;11/20/1967;4 Debs Court;9 745 586 157 915;dabelson1@gmail.com;2		
3	3;Merrily;Cain;12/28/1945;77713 Glendale		

Гаразд. Тепер перевір якість даних в прикріпленому файлі відповідно до стандарту ISO 25012

Рис.3.109 Запит в чату gpt на перевірку якості даних

Результат перевірки при такому запиті виявився цілком очікуваним, і модель змогла виявити лише 2 помилки (рис.3.110).

Missing Data Analysis		
		0
1	ID	0
2	First_name	0
3	Last_name	0
4	Date of birth	0
5	Address	0
6	Phone Number	0
7	Email	0
8	Connection type	0
9	Services	1
10	Tariff plan	0
11	Account Status	1

Рис.3.110 Вивід виявлення двох помилок

Даними помилками є пропуски значень, присутні в датасеті (рис.3.111).

1. Повнота (Completeness):	
•	Всі стовпці, окрім стовпців <code>`Services`</code> та <code>`Account Status`</code> , не містять пропусків.
•	Пропуски у даних:
•	<code>`Services`</code> : 1 пропуск
•	<code>`Account Status`</code> : 1 пропуск

Рис.3.111 Виявлені помилки

Наступним кроком є більш конкретизований запит. В ньому уточнюється, по чому можна провести перевірку номерів, дані з якими датами народження,

вважаються актуальними та достовірними, а також уточнення по бажаному виводу результатів (рис.3.112).

Уявіть, що ви - інструмент для перевірки даних. Вам потрібно перевірити якість даних відповідно до стандарту ISO 25012 у файлах, які я вам надсилаю. Ваше завдання - проаналізувати мій файл і знайти якомога більше даних, які не відповідають критеріям якості даних ISO 25012. За можливості, ви повинні перевірити контекст даних, беручи до уваги, до якої сфери вони належать. Що стосується номерів телефонів, ви можете перевірити їхню достовірність за довжиною полів, оскільки можлива довжина телефонного номера відома. Для дат народження перевірте їх коректність на основі сьогодення дати та врахуйте можливу тривалість життя людини. Також не повинно бути нульових значень. Якщо можливо, перевірте, чи не є значення конкретного поля частиною списку можливих значень і чи немає значень, які не входять до цього списку. За результатами аналізу виведіть таблицю з усіма знайденими недостовірними даними, а в окремій таблиці надайте статистику про те, скільки і яких саме порушень ви знайшли.

Рис.3.112 Конкретизований запит чату gpt

Після даних уточнень модель змогла виявити більше помилок, але також значно збільшилась кількість хибних розпізнавань (рис.3.113).

Статистика порушень	
Issue	Count
Invalid connection type	1000
Invalid account status	712
Missing values	2
Invalid date of birth	2
Invalid phone number	2

Рис.3.113 Вивід помилок після конкретизованого запиту

Для зменшення хибних спрацьовувань на рядках Connection type та Account status, так як ці поля мають список можливих значень, надаю цей список моделі (рис.3.114)

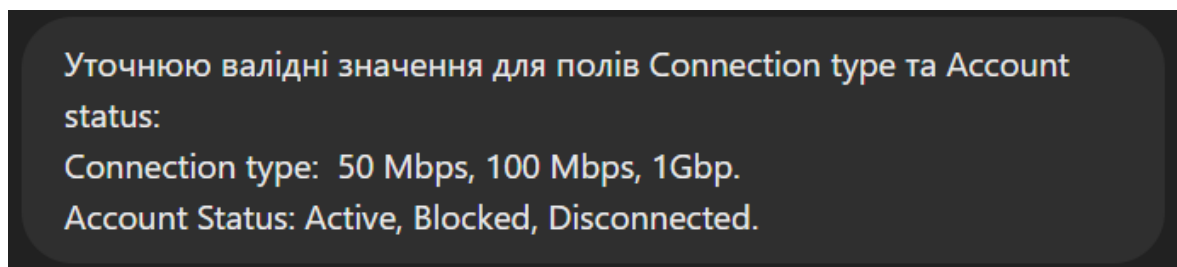


Рис.3.114 Уточнення значень для полів Connection Type та Account Status

Після такого уточнення кількість хибних спрацьовувань значно зменшилась, але кількість хибних спрацьовувань для поля Connection type досі велика (рис.3.115)

Статистика порушень	
Issue	Count
Invalid connection type	373
Invalid account status	3
Missing values	2
Invalid date of birth	2
Invalid phone number	2

Рис.3.115 Вивід хибних спрацьовувань для поля Connection Type

Знову уточнюю можливі значення для поля Connection type (рис.3.116)

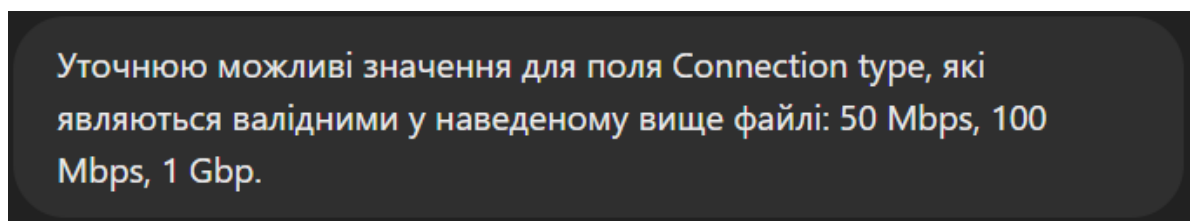


Рис.3.116 Уточнення можливих значень для Connection Type

Як результат, після повторного уточнення кількість хибних спрацьовувань значно зменшилась (рис.3.117)

Статистика порушень	
Issue	Count
Invalid account status	3
Missing values	2
Invalid date of birth	2
Invalid phone number	2
Invalid connection type	2

Рис.3.117 Вивід статистики порушень після повторного уточнення

Невалідні дані, які вдалось виявити моделі зображені нижче (рис.3.118)

Index	Issue	Invalid Data
17	Missing values	{'ID': 18, 'First_name': 'Husein', 'Last_name': 'Abd', 'Date of birth': '09/20/1983', 'Address': '22 Roth Center', 'Phone Number': '824 123 5678', 'Email': 'husein.abd@example.com', 'Connection type': '1 Gbp', 'Services': nan, 'Tariff plan': 'Basic', 'Account Status': nan}
53	Missing values	{'ID': 54, 'First_name': 'Cecelia', 'Last_name': 'Abelson', 'Date of birth': '11/20/1967', 'Address': '4 Debs Court', 'Phone Number': '9 745 584 123', 'Email': 'cecelia.abelson@example.com', 'Connection type': '100 Mbps', 'Services': nan, 'Tariff plan': 'Premium', 'Account Status': 'Active'}
67	Invalid date of birth	07.05.2050
160	Invalid date of birth	01.06.1900
342	Invalid phone number	502 914
437	Invalid phone number	3 756 842 281 421 000
508	Invalid connection type	one Gbp
640	Invalid connection type	one hundred Mbps
38	Invalid account status	Активний
163	Invalid account status	Блокировка

Рис.3.118 Вивід виявлених невалідних значень

Хоча модель правильно виявила, в яких рядках присутні невалідні дані, вона правильно виводить дані для рядків 17 та 53.

Після цього уточнюю можливі значення для полів Services та Tariff Plan для підвищення ефективності моделі (рис.3.119)

Уточнюю валідні значення для полів Services та Tariff plan:
Services: Phone, Internet, Cable TV.
Tariff plan: Basic Plan, Gold Plan, Premium Package.

Рис.3.119 Уточнення для полів Services та Tariff Plan

Знайдені помилки відображені нижче (рис.3.120).

Row Index	Issue	Invalid Data
17	Missing values	{'ID': 18, 'First_name': 'Husein', 'Last_name': 'Abd', 'Date of birth': '09/20/1983', 'Address': '22 Roth Center', 'Phone Number': '824 123 5678', 'Email': 'husein.abd@example.com', 'Connection type': '1 Gbp', 'Services': nan, 'Tariff plan': 'Basic Plan', 'Account Status': nan}
53	Missing values	{'ID': 54, 'First_name': 'Cecelia', 'Last_name': 'Abelson', 'Date of birth': '11/20/1967', 'Address': '4 Debs Court', 'Phone Number': '9 745 584 123', 'Email': 'cecelia.abelson@example.com', 'Connection type': '100 Mbps', 'Services': nan, 'Tariff plan': 'Premium Package', 'Account Status': 'Active'}
67	Invalid date of birth	07.05.2050
160	Invalid date of birth	01.06.1900
342	Invalid phone number	502 914
437	Invalid phone number	3 756 842 281 421 000
508	Invalid connection type	one Gbp
640	Invalid connection type	one hundred Mbps
17	Invalid account status	NaN
38	Invalid account status	Активний
163	Invalid account status	Блокировка
8	Invalid services	Smartphone
292	Invalid tariff plan	Minimum Plan

Рис.3.121 Знайдені помилки після уточнень

Як результат, моделі штучного інтелекту вдалось виявити лише 12 помилок з 14. Їй не вдалось виявити невалідні дані в двох випадках: неправильна електронна пошта та неправильний реєстр літер в імені. В цілому можна казати про те, що штучний інтелект може аналізувати дані, але для такого аналізу потрібно вказувати велику кількість уточнень, які він не відразу сприймає: мені довелося двічі вказати дозволені значення для певних полів, щоб він вірно визначив помилки. Наразі довіри до штучного інтелекту при аналізі даних немає, і надійніше валідувати дані через спеціальні інструменти, так як при їх використанні користувач може бути впевнений в тому, що не потрібно вказувати певні критерії для валідації даних двічі.

Висновки

Будь-яка система не може існувати без даних. Дані є основою всього, що нас оточує. Проте, важливо дотримуватись того, щоб дані були валідними. Для валідації даних розробляється безліч інструментів, але під час їх розробки та тестування важливо перевіряти їхню роботу на певних категоріях даних. Тому важливою частиною роботи з інструментами є підготовка датасетів. Датасети повинні відображати реальні сценарії використання даних у сфері телекомунікацій. Основними критеріями таких даних є точність і відповідність, але для того, щоб бути впевненими в роботі інструментів і знати наскільки ефективно вони справляються із поставленим завданням краще формувати помилки вручну. Створення штучних помилок дозволяє оцінити роботу інструментів на прикладі того, скільки помилок із заданих він знайшов. Таким чином тестування інструментів однаковими даними і в однакових умовах дозволяє успішно оцінити їх і об'єктивно порівняти між собою.

РОЗДІЛ 4

РЕЗУЛЬТАТИ РОБОТИ ІНСТРУМЕНТІВ

4.1 Оцінка роботи інструментів зі згенерованими даними

Порівняння вибраних інструментів і вибір найкращого з них є головною метою даного дослідження. Тому у цьому розділі ми підсумуємо оцінки протестованих інструментів, включаючи Data Ladder, Alteryx та Informatica Data Quality за вибраними критеріями.

У подальшому міра відповідності вибраним критеріям буде оцінюватись від 1 до 3 балів, де 1 - незадовільна оцінка (інструмент взагалі не відповідає даному критерію), 2 - середня (даний інструмент задовольняє даний критерій, але не ідеально), 3 - відмінна (інструмент повністю відповідає даному критерію), 0 - інформація по даному критерію відсутня, або у інструмента відсутній функціонал, який покриває даний критерій.

Отже, першим протестованим інструментом був Data Ladder

Даний інструмент має дуже простий інтерфейс через свій обмежений функціонал, попри це, деякі елементи інтерфейсу виявились незрозумілими, тому по критерію зручності використання він отримав 2 бали.

Інструмент має плавний інтерфейс, який не має мікрозатримок, а також інтерфейс чітко реагує на взаємодію користувача з ним, тому за критерій часу відгуку інструмент отримує 3 з 3.

Після реєстрації для пробного періоду, я регулярно отримувала листи з запитаннями щодо того, як проходить тестування інструменту, і що в разі потреби я можу звернутись в службу підтримки. Також при зверненні в службу підтримки я швидко отримала відповідь, тому по критерію Help and Support інструмент отримує 3 бали з 3.

Так як даний інструмент являється нативним додатком який може працювати без підключення до інтернету, за критерієм Offline tool я оцінила інструмент на 3 з 3.

Пробний період який доступний після реєстрації складає 30 днів, чого достатньо для тестування даного інструменту, тому за критерієм Trial period я оцінюю інструмент на 3 з 3.

Щодо ціни, то її можна дізнатись лише після зв'язку із менеджером по продажах, але так як відповіді від менеджера з продажу не було отримано, то цінова категорія даного інструменту залишається невідомою. За це ставлю 0 балів.

Інструмент не має функцій, які б підвищували його інклюзивність, але завдяки тому, що інтерфейс складається з друкованого тексту, програми розпізнавання тексту зможуть зчитувати з додатка інформацію, тому за критерієм Accessibility Features я оцінюю даний інструмент на 1 бал.

Даний інструмент не має жодної незвичайної функціональності, тому за критерій Impressive Feature інструмент отримує 0 балів.

Критерій Test suite preconditions (single login) для даного інструменту я оцінила в 3 бали, так як даний додаток не потребує авторизації.

Даний інструмент вміє працювати як з малими об'ємами даних, так і з великими, тому за критерій Scalability інструмент отримує 3 бали.

Так як даний додаток підтримує можна вважати що один тестовий сценарій і не має жодних автоматизованих тестів у пробній версії, але в платній версії вони присутні, я оцінила даний інструмент за критерієм DDT (Data-Driven Testing) в 1 бал.

За критерієм Variety of Test Cases даний інструмент я оцінила в 1 бал, так як даний інструмент має один тестовий сценарій, який не можна змінювати.

По критерію Data Complexity Handling я оцінила даний інструмент в 3 бали, так як він вміє працювати із складними структурами даних а також адаптивних до динамічних змін.

За критерієм Data Integrity Verification я оцінюю даний додаток в 2 бали з 3, так як цей додаток перевіряє цілісність даних при виконанні тестування цих даних.

Test creation time я оцінила в 1 бал, так як в даному додатку можливий лише один тест, який вже попередньо створений, і користувач може лише змінювати параметри для даного тесту.

Test execution time я оцінила в 2 бали, так як порівняно з іншими інструментами, в даному інструменті виконання тесту триває довше.

Test report я оцінила в 2 бали, так як при виконанні тесту створюється звітність про результати даного тесту, але не будуються звіти про те, які зміни були зроблені, які дані зазнавали змін, який користувач проводив зміни і так далі.

Даний інструмент має можливість підключатись та взаємодіяти із різними СУБД, але їхня кількість менша ніж в конкуруючих інструментах, тому за критерій DB integration інструмент отримує 2 бали.

Даний інструмент може взаємодіяти з різними типами API, але налаштування такої взаємодії складніша ніж в інструментах конкурентах, тому я оцінила критерій API Integration в 2 бали.

За критерій Customization and extensibility я поставила інструменту 1 бал, так як даний інструмент надає базові можливості налаштування додатка.

За критерієм SDK я оцінила даний інструмент в 2 бали з 3, так як даний інструмент дає програмний розробницький набір, достатній для того щоб дозволити вашому додатку надавати доступ на взаємодії з інструмента валідації даних.

За критерієм CI/CD integrations я поставила даному додатку 1 бал, так як він може інтегруватися з малою кількістю сервісів для процесів безперервної інтеграції/безперервного розгортання.

За критерій Data cleansing and transformation я поставила інструменту 2 бали, так як він має функції для очищення та трансформації даних, але кількість функцій менша ніж в конкурентів.

Інструмент вміє візуалізовувати певні кроки аналізу даних та зберігати дані візуалізації в окремий звіт, тому за критерієм Data visualization я оцінила цей додаток на 2 бали з 3.

Наступним протестованим інструментом став Alteryx.

Usability даного інструменту я оцінила в 3 бали, зокрема, через те, що його інтерфейс доступний і зрозумілий навіть для людей без якихось спеціальних технічних знань. Тобто зайшовши в інструмент можна досить легко і швидко в ньому розібратись.

Наступний критерій, а саме час реакції, я також оцінила в 3 бали, адже час реакції інструменту на дії користувача досить швидкий, швидко обробляються натискання, інтерфейс інструменту не зависає.

Help and Support даної утиліти теж на максимум, адже я отримала швидко відповідь на пошту і без проблем змогла отримати доступ до пробного періоду, навіть враховуючи ті фактори, що я реєструвалась як студент і потрібно було перевірити наявність студентського, все одно, реєстрація не зайняла більше 5 хвилин часу.

Offline tool - 3 бали за те що Alteryx являється нативним додатком який підтримує основні платформи, такі як Linux, MAC та Windows, а також який може працювати без підключення до інтернету.

Пробний період я оцінила у 2 з 3, адже для студентів та науковців інструмент надає цілих 365 днів пробного періоду, але для звичайних користувачів отримати доступ до пробного періоду можливості немає.

Ціна також на 2 з 3, тому що хоч і ціна найбільша з усіх протестованих утиліт, інструмент виправдовує її своїми функціональними можливостями.

Accessibility Features - 3. Alteryx надає кілька функцій доступності, щоб забезпечити можливість використання платформи людьми з обмеженими

можливостями. Сюди входять такі функції як, сумісність з пристроями для читання екрану, висококонтрастні режими, змінні розміри тексту та інших елементів, альтернативний текст для зображень та іконок.

Impressive Feature я оцінила 3 балами, адже це єдиний інструмент з протестованих, в якому я помітила дійсно особливу функцію - машинне навчання на основі даних.

Test suite preconditions (single login) – максимальний бал, адже після встановлення програми потрібно лише один раз підтвердити свій логін та ключ ліцензії, і після цього для виконання тестів логінитись вже не потрібно.

Scalability також 3, адже інструмент ефективно опрацьовує як малі так і великі обсяги даних, що є безперечним плюсом для сфери телекомунікацій.

DDT (Data-Driven Testing) оцінила в 3, адже одну блок-схему можна адаптувати під різні джерела інформації, просто змінивши саме джерело.

Variety of Test Cases я оцінила в 3 бали з 3, так як завдяки своєму великому функціоналу, інструмент може перекривати величезну кількість тестових кейсів.

Data Complexity Handling на 3 з 3, так як інструмент вміє працювати з складними структурами даних, а також вміє адаптовуватись до джерел даних з динамічною структурою.

Data Integrity Verification знову ж таки 3 бали, так як інструмент перевіряє цілісність та достовірність даних під час кожного тесту, та в разі проблем сповіщає користувача.

Test creation time оцінила на 3 з 3, через те, що блоки розміщуються через графічний інтерфейс, і структуру тесту наочно видно, тому створення тесту не займає багато часу.

Test execution time - час виконання тестів залежить напряду від потужності пристрою, на якому виконуються тести. Так як на моєму не дуже потужному пристрої тести виконувались швидко, я оцінила даний інструмент за цим критерієм на 3 бали з 3

Test report - оцінила даний критерій на 3 з 3, тому що інструмент виводить інформацію під час проведення тестів про помилки та попередження, також після виконання тесту візуалізує результат.

DB integration - є можливість інтеграції з великою кількістю СУБД, також можна інтегрувати з різними хмарними сховищами, тому 3 бали.

API integration - даний інструмент може взаємодіяти з API різних типів, але свого API даний інструмент не має, тому я оцінюю його на 2 бали з 3.

Customization and extensibility - утиліта дає можливість користувачу адаптовувати інструмент під себе і встановлювати додаткові розширення, тому 3 з 3.

SDK - так як розробник надає SDK і він надає широкі можливості для покращення інструменту, то оцінила даний критерій на 3 з 3.

CI/CD integration - 3 бали, адже інструмент має можливість інтеграції з різними системами безперервної інтеграції/безперервного розгортання.

Data cleansing and transformation - інструмент надає широкий функціонал для очищення і трансформації даних на рівні з іншими конкурентоспроможними утилітами, тому оцінила якість інструменту по даному критерію на 3 з 3.

Щодо Data visualization, то інструмент загалом добре візуалізує дані, профільовані дані інструмент відображає у вигляді інформаційної панелі, діаграм та графіків, але даний інструмент трішки поступається Informatica Data Quality в даному критерії, відсутні дані про патерни для кожного поля, тому оцінила на 2 з 3.

Наступний протестований інструмент, це Informatica Data Quality. Інструмент має зручний інтерфейс, хоча функціональність деяких елементів з першого погляду є незрозумілою. При подальшому використанні користувач звикає до інтерфейсу і може комфортно використовувати інструмент. Також будь-яка помилка при використанні інструменту детально описується, завдяки

чому можна швидко знайти рішення. Тому я оцінила зручність використання на 2 бали з 3.

Даний інструмент має малий час відповіді на дії користувача, сторінки інструменту завантажуються швидко, але в період високого навантаження час завантаження сторінок може збільшитись, але не критично. Також відсутні мікрозатримки при взаємодії з інтерфейсом. В цілому по цьому критерію інструмент добре себе проявив, але через затримки під час високого навантаження на сервер я оцінила даний інструмент на 2 бали з 3.

Наступний критерій стосується підтримки та допомоги. Даний інструмент має різні способи зв'язку із технічною підтримкою, включаючи чат-бот, електронна пошта або форум. Служба підтримки швидко реагує на звернення. Також інструмент має велику кількість документації та детальних інструкцій, тому за цим критерієм він безумовно заслуговує на 3 бали.

Так як даний інструмент є онлайн сервісом, то користувач не має змоги працювати з ним без підключення до інтернету, тому за за можливість використання без інтернету даний сервіс отримує 0 балів.

Даний інструмент дозволяє легко отримати доступ до пробної версії шляхом простої реєстрації, при цьому через пошту будь-якого домену. Пробний період триває 30 днів, чого повністю достатньо для того щоб оцінити функціональні можливості даного інструменту. Тому за критерієм Trial period інструмент отримує 3 бали з 3.

Щодо ціни, то сервіс має конкурентну ціну. Функціонал даного інструменту настільки великий та різноманітний, що повністю виправдовує ціну. Саме завдяки цьому за цим критерієм я оцінила даний інструмент на 3 бали.

Даний інструмент не має жодних додаткових функцій для забезпечення інклюзивності, і єдиним плюсом є те, що інтерфейс складається в основному з друкованого тексту, а не картинок, тому завдяки цього програми зчитувачі

тексту зможуть без проблем зчитувати інформацію з сайту, і саме через цей критерій інструмент отримав 1 бал.

Щодо вражаючих функцій, то даний інструмент виділяється зручним сервісом для моніторингу загального стану інструменту: стан робочих оточень, стан серверів сервіса, стан запущених завдань на виконання та інші важливі дані, які потрібно моніторити для забезпечення працездатності інструменту в конкретного користувача. За цю функцію інструмент отримав 2 бали.

Наступним критерієм є *test suite preconditions (single login)*. За цим критерієм інструмент отримує від мене 3 бали з 3, так як він потребує залогінитись в профіль лише раз за сесію, і під час виконання тестів авторизація не вимагається.

Щодо масштабованості, то цей інструмент також отримує 3 бали, так як він чудово працює як з невеликою кількістю даних, так і з дуже великою.

За критерієм *DDT* інструмент також отримує 3 бали через свою гнучкість: тести які створені в даному інструменті можуть обробляти дані з різних наборів вхідних даних без змін, або з мінімальними змінами.

Даний інструмент містить велику кількість різноманітних функцій, завдяки яким можна створити тести для безлічі сценаріїв, тому все обмежується лише уявою користувача. Через таку можливість інструмента покривати велику кількість різних тестових кейсів та сценаріїв, інструмент отримує 3 бали з 3 за критерієм *Variety of Test Cases*.

Інструмент вміє обробляти складні структури даних, а також адаптивний до змін в структурі даних. Він працює з даними тільки в разі вдалого їхнього розпізнавання, що йому вдається в більшості випадків. Тому за критерієм *Data Complexity Handling* інструмент отримує 3 бали з 3.

Перед зберіганням даних після їхнього завантаження інструмент проводить перевірку на правильність даних, і в разі проблем він сповіщає користувача, то сталась певна помилка. Також при використанні даних в тестуванні, першим кроком який виконує інструмент це перевірка цілісності

даних, також після обробки інструмент знову ж таки перевіряє цілісність даних. На жаль, інструмент не зберігає чіткої історії змін в даних. Тому за критерієм Data Integrity Verification інструмент отримує 2 бали з 3.

Створення тестів у даному інструменті спочатку відбувається довго через відсутність певних стандартних шаблонів для об'єктів, і через це доводиться створювати все з 0. Після того як в користувача назбирається власна колекція об'єктів, то створення тестів відбувається набагато швидше та зручніше, тому по критерію Test creation time інструмент отримує 2 бали з 3.

Щодо часу виконання тестів, то він дещо більший ніж в інших протестованих інструментах. Це обумовлено тим, що сервіс обробляє дані в середовищі виконання, яке налаштовує користувач, тому швидкість виконання тестів залежить як від потужності пристрою, на якому розміщене робоче оточення, так і від швидкості інтернету, а також геоположення. Через це інструмент по критерію Test execution time отримує 2 бали з 3.

Після виконання аналізу або тестування даних, інструмент у зручному вигляді звітує про результати виконаних тестів. Інструмент надає зрозумілі графіки та діаграми. За критерієм Test report, інструмент отримує 3 бали.

Критерій DB integration оцінює можливість інструменту підключатись до джерел різних типів баз даних. Даний інструмент має велику кількість шаблонів до підключення до баз даних, а також підтримує підключення через ODBC. Тому за цим критерієм я оцінила даний інструмент на 3 бали з 3.

Цей інструмент вміє працювати з різними протоколами API, і дозволяє легко налаштувати таке з'єднання, а також з цим інструментом можна також взаємодіяти через API, тому за цим критерієм я оцінила цей інструмент на 3 з 3.

Інструмент можна гнучко налаштовувати а також розширяти за допомогою розширень із магазину розширень, завдяки чому користувачі за потреби можуть встановити необхідні для них розширення, щоб адаптуватись під певні бізнес-вимоги, а тому за критерієм Customization and extensibility я оцінюю даний інструмент на 3 з 3.

Розробники даного інструменту надають SDK для того, щоб інші могли реалізувати підтримку свого інструменту з боку informatica cloud, але на цьому можливість даного SDK закінчується, тому критерій наявності SDK 2 з 3.

Даний інструмент легко можна інтегрувати з процесами безперервної інтеграції/безперервного розгортання, за що по критерію CI/CD integrations інструмент отримує 3 бали

Даний інструмент має багато функцій для очищення та різного роду трансформацій даних, тому за критерієм Data cleansing and transformation отримує 3 бали з 3.

Після виконання тестів або профілювання даних, інструмент надає результати у зручному візуалізованому вигляді, який зручно переглядати та аналізувати, тому за критерієм Data visualization інструмент отримує 3 бали з 3.

Нижче наведена візуальна таблиця порівняння протестованих інструментів і загальні результати (рис.4.1). Як можна побачити, найкращим з протестованих інструментів стає Alteryx, зібравши найбільшу кількість балів, цей інструмент дійсно вражає своїми функціональними можливостями. Друге місце з невеликим відривом посідає Informatica Data Quality і останнє місце Data Ladder.

Tools	Data Ladder	Alteryx	Informatica Data Quality
Personal impression			
Usability	2	3	2
Response Time	3	3	2
Help and Support	3	3	3
Accessibility			
Offline tool	3	3	0
Trial period, days	3	2	3
Price	0	2	3
Accessibility Features	1	3	1
Impressive Feature	0	3	2
Test Optimizations			
Test suite preconditions (single login)	3	3	3
Scalability	3	3	3
DDT	1	3	3
Tests			
Variety of Test Cases	1	3	3
Data Complexity Handling	3	3	3
Data Integrity Verification	2	3	2
Tests Execution and Reporting			
Test creation time	1	3	2
Test execution time	2	3	2
Test report	2	3	3
Capabilities			
DB integration	2	3	3
API integration	1	2	3
Customization and extensibility	1	3	3
SDK	2	3	2
CI/CD integrations	1	3	3
Data cleansing and transformation	2	3	3
Data visualization	2	2	3
Total			
	44	68	60

Рис.4.1 Порівняльна таблиця протестованих інструментів

4.2 Висновки з експериментального аналізу

Провівши експериментальний аналіз трьох інструментів для валідації даних - Data Ladder, Alteryx та Informatica Data Quality, я готова зробити певні висновки.

Alteryx це потужний і повнофункціональний інструмент, який чудово підходить для великих компаній з великими обсягами даних та складними вимогами до забезпечення якості даних. Він має широкий спектр функцій для профілювання, очищення, трансформувannya даних, а також високу швидкість виконання операцій. Однак, даний інструмент має складний рівень входження через незвичайний інтерфейс та велику кількість функцій, тому для ефективного і результативного використання доведеться витратити досить багато часу на навчання та розуміння даного інструменту.

Data Ladder, навпаки, є простим і зрозумілим інструментом, орієнтованим на малий бізнес та разові перевірки даних. Він забезпечує легкість використання, автоматичне розпізнавання типів полів та ефективний пошук дублікатів. Проте, його простота обмежує функціональність, і він не здатен до автоматизації та складних операцій з даними, присутніх в інших інструментах.

Informatica Data Quality це гнучкий і потужний інструмент для середніх та великих організацій. Він поєднує зручний інтерфейс з широким спектром функцій для профілювання, очищення, інтеграції даних та можливістю інтеграції з іншими системами. Як і Alteryx, він вимагає певного часу на освоєння, але це виправдано його можливостями.

Загалом, вибір інструменту для валідації даних залежить від конкретних потреб та ресурсів організації. Alteryx та Talend Data Quality є найкращими варіантами для великих компаній з великими обсягами даних та складними вимогами до якості даних, тоді як Data Ladder більше підходить для малого бізнесу або одноразових перевірок. Врахування переваг та недоліків кожного

інструменту допоможе зробити правильний вибір для забезпечення ефективної валідації даних.

ЗАГАЛЬНІ ВИСНОВКИ ПО РОБОТІ

В результаті виконання дипломної роботи було розглянуто і описано ряд популярних інструментів для владіції даних. Серед них отримано доступ і протестовано три: Data Ladder, Alteryx та Informatica Data Quality. Тестування продемонструвало можливості, переваги та недоліки кожного з них, дозволило отримати загальну картину про функціонал і порівняти їх між собою.

У висновку, Data Ladder посідає третє місце, маючи досить простий функціонал і обмежені функції, орієнтований даний інструмент на тестування у малих компаніях та з невеликими обсягами даних. Друге місце займає Informatica Data Quality, маючи загальну кількість 60 балів по мірі відповідності вибраним критеріям, це потужний і гнучкий інструмент, який зміг справитись із поставленим завданням і продемонстрував свій широкий функціонал. Даний інструмент повністю підходить для середніх і великих організацій. Перше місце заслужено посідає Alteryx через свої потужні і повнофункціональні можливості. Даний інструмент чудово підходить для масштабованих організацій із великими обсягами даних та складними взаємодіями. Крім практичного застосування інструментів наведено факти про важливість та проблематику валідації даних в сучасному світі. Також на практиці дослідження ознайомлює і дозволяє більше дізнатись про стандарт якості ISO 25012, його основні критерії оцінки та роль у забезпеченні якісних даних. Використання стандарту ISO 25012 як основи для оцінки інструментів сприяє забезпеченню високої якості даних, що є запорукою успішної діяльності телекомунікаційних компаній у сучасних умовах.

Щодо складеної методики тестування, то вона повністю підтвердила очікування і виявилась дієвою для аналізу інструментів. Також дослідження дозволяє ознайомитись із процесом генерації даних, створенням датасетів та прикладом їх застосування.

Загалом виконана робота може стати чудовим гайдом для ознайомлення та користування протестованими інструментами, надаючи детальні інструкції щодо вибору та впровадження інструментів для валідації даних. Дане дослідження може бути корисним з боку рекомендацій для організацій з різними потребами і різного масштабу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. admin. ETL или ELT: какой процесс работы с данными дает оптимальный результат - RBC group. *RBC group*. URL: <https://www.rbcgrp.com/ua/etl-ili-elt-kakoj-process-raboty-s-dannymi-daet-optimalnyj-rezultat/>
2. Остапов О. Стандарт якості даних ПЗ – детально розбираємо ISO 25012. *DOU*. 2023. URL: <https://dou.ua/forums/topic/44154/>
3. Валідація. Тестування валідації. *Онлайн-курси від компанії QATestLab|Головна сторінка*. URL: <https://training.qatestlab.com/blog/technical-articles/validation-testing/>
4. ISO - International Organization for Standardization. URL: <https://www.iso.org/obp/ui/ru/#iso:std:iso-iec:25012:ed-1:v1:en/>
5. ДСТУ ISO/IEC 25012:2016 Інженерія систем і програмних засобів. Вимоги до якості систем і програмних засобів та її оцінювання (SQuaRE). Модель якості даних (ISO/IEC 25012:2008, IDT).
6. Top 7 Data Validation Tools to Explore in 2024 | Astera. *Astera*. URL: <https://www.astera.com/type/blog/data-validation-tools/>
7. OpenRefine. OpenRefine. URL: <https://openrefine.org/>
8. ДСТУ ISO/IEC 25010:2016 Інженерія систем і програмних засобів. Вимоги до якості систем і програмних засобів та її оцінювання (SQuaRE). Моделі якості системи та програмних засобів (ISO/IEC 25010:2011, IDT)
9. 11 найкращих генераторів даних MS SQL (2024).. *Data Recovery Blog*. URL: <https://www.datanumen.com/uk/блоги/11-найкращих-генераторів-даних-ms-sql-завантажити-безкоштовно/>
10. Mockaroo - Random Data Generator and API Mocking Tool | JSON / CSV / SQL / Excel. URL: <https://www.mockaroo.com>

11. Остапов О. Інструменти для тестування даних, codeless-автоматизація та створення mock-сервісу – три дипломи моїх студентів. *DOU*. 2024. URL: <https://dou.ua/forums/topic/48305/>