Санкт-Петербургский государственный университет

Кафедра информационно-аналитических систем

Группа 21.Б08-мм

# Примитивы профилирования данных в Desbordante: обзор и сравнение существующих инструментов

***ГОНЧАРОВ Даниил Юрьевич***

Отчёт по учебной практике
в форме «Сравнение»

Научный руководитель:
ассистент кафедры ИАС Г. А. Чернышев

Санкт-Петербург
2024

# Оглавление

# Введение

В современном мире данные стали фундаментальным ресурсом, формирующим экономику и повседневные процессы. Умение эффективно собирать, обрабатывать и анализировать информацию является залогом конкурентоспособности и успеха, поэтому нельзя переоценить важность специализированных приложений для работы с данными. Подобные инструменты фокусируются на задачах, связанных с так называемым качеством данных.

Одной из таких задач является сбор данных о самих данных — профилирование. Часть приложений предлагает понятную интерпретацию и красивую визуализацию выявленной информации, другие на основе полученных метаданных помогают определить возможные ошибки и пропуски, улучшая и насыщая таким образом первоначальные данные.

Само профилирование в свою очередь можно условно разделить на простое и наукоёмкое. К простому относится вся статистическая информация, включающая в себя данные о колонках, типах и возможных шаблонах, а более трудоёмкое выявление всевозможных зависимостей и закономерностей в данных можно считать наукоёмким профилированием.

Во время старта проекта Desbordante проводились обзоры приложений-профилировщиков на способность проведения анализа более сложного чем статистического. В них было показано, что инструментов с подобным функционалом крайне мало. Однако, так как с того момента прошло много времени, было принято решение обновить информацию о статусе наукоёмкого профилирования в приложениях по работе с данными. Для этого необходимо провести обзор и сравнение существующих инструментов на основе примитивов, реализованных в Desbordante. Это и стало темой настоящей работы.

# 1 Постановка задачи

Цель данной работы заключается в поиске и сравнении инструментов, возможно подходящих для наукоёмкого профилирования. Для выполнения этой цели были поставлены следующие задачи:

1. Провести обзор существующих исследований в сфере работы с данными.
2. Отобрать подходящие для сравнения инструменты.
3. Провести обзор отобранных инструментов.
4. Провести сравнение отобранных инструментов.
5. Сделать выводы о статусе наукоёмкого профилирования среди отобранных инструментов.

# 2 Обзор существующих обзоров

В данной главе рассматриваются статьи, в которых в том или ином виде проводился обзор инструментов для работы с данными.

## 2.1 A Survey of Data Quality Measurement and Monitoring Tools

Данное исследование [5] направлено на обзор приложений для мониторинга и измерения качества данных. Авторами было найдено 667 инструментов, из которых в последствии было отобрано 13. Отобранные инструменты подверглись детальному анализу на возможность профилирования данных, измерения качества данных и мониторинга качества данных.

В результате было выявлено, что базовые функции профилирования поддерживаются большинством инструментов, однако, более сложные функции, связанные с поиском зависимостей и многоколоночным профилированием реализованы лишь некоторыми инструментами.

Также были сделаны следующие выводы:

1. Нет ни одного инструмента, который бы реализовывал широкий спектр метрик для измерения качества данных;

2. В сфере профилирования данных и измерении качества данных необходимо больше автоматизации.

## 2.2 Evaluation of freely available data profiling tools for health data research application: a functional evaluation review

Это исследование [6] было направлено на оценку возможностей профилирования данных при помощи бесплатных приложений. Авторы поставили задачу выявить инструменты с широкой функциональностью

и высокой производительностью, поскольку здравоохранительные организации зачастую имеют ограниченные возможности и недостаточный опыт в области обработки данных.

Изначально было выбрано 28 инструментов, однако после ознакомления с документацией было решено сравнить 8, которые имели наибольший потенциал.

В результате проведённого тестирования был сделан вывод, что по крайней мере два инструмента показали высокую производительность на рассматриваемых статистиках[1] и могут использоваться здравоохранительными организациями, однако выбор подходящего инструмента всё равно должен основываться на инфраструктуре организации и опыте в области обработки данных и программирования.

## 2.3 A Survey on Data Quality Dimensions and Tools for Machine Learning

Данное исследование [16] рассматривает важность измерения качества данных при работе с инструментами для машинного обучения. В ходе работы обсуждаются различные измерения качества данных и их влияние на модели машинного обучения. Рассматривается 17 инструментов, вышедших за последние 5 лет и сравнивается их функционал.

В конце авторы приводят своё видение на дальнейшую разработку приложений для измерения качества данных. Основная идея заключается в том, что весь процесс от загрузки данных до выдачи результатов и отчетов должен проходить через одно приложение, которое должно поддерживать автоматический мониторинг качества данных.

Выводы, сделанные авторами, гласят, что интеграция искусственного интеллекта для решений задач связанных с качеством данных уже вошло в практику в современных приложениях и только будет развиваться в дальнейшем.

[1]Стоит отметить, что во второй таблице данного исследования допущена неточность, так как инструмент pandas profiling (нынешний YData [20]) не имеет некоторого приписанного ему функционала, а именно поиска зависимостей и части пунктов из многоколоночного профилирования.

## 2.4 From Cleaning before ML to Cleaning for ML

В данной работе [9] рассматривается подход к очистке данных в приложениях машинного обучения, который предполагает очистку на протяжении всего жизненного цикла ML-приложения.

Авторы рассматривают устоявшийся подход к процессу работы приложений машинного обучения, который разделен на три этапа (подготовка, разработка, развёртывание). Параллельно с этим рассматриваются всевозможные инструменты и подходы, используемые для подготовки и очистки данных.

В ходе экспериментов выясняется, что очистка данных может как улучшать, так и ухудшать производительность модели, поэтому предлагается архитектура, которая интегрирует очистку данных в весь жизненный цикл, учитывая потребности каждой фазы.

## 2.5 Итоги

Далее представлена таблица с информацией по разобранным исследованиям, а также в краткой форме выделены важные моменты каждого из них. В конце будет подведён итог о теме данной работы в рассмотренных исследованиях.

Таблица 1: Исследования

| Статья | Год | Фокус | Место публикации | Объём |
|---|---|---|---|---|
| [5] | 2022 | Измерение и мониторинг качества данных | Frontiers in Big Data | 30 стр. |
| [6] | 2022 | Профилирование данных в сфере здравоохранения | BMJ Open | 12 стр. |
| [16] | 2024 | Машинное обучение и качество данных | IEEE AITest 2024 | 12 стр. |
| [9] | 2021 | Машинное обучение и подготовка данных | IEEE Data Engineering Bulletin | 18 стр. |

1. **A Survey of Data Quality Measurement and Monitoring Tools**

- Лишь малая часть инструментов реализует более сложное профилирование.
- В сфере профилирования необходимо больше автоматизации.

2. **Evaluation of freely available data profiling tools for health data research application: a functional evaluation review**

   - Инструменты для профилирования упрощают работу в организациях здравоохранения.
   - Из рассмотренных лишь два инструмента продемонстрировали достойный функционал и производительность.

3. **A Survey on Data Quality Dimensions and Tools for Machine Learning**

   - Измерение качества данных очень важно при машинном обучении.
   - Существует тенденция на внедрение ИИ в задачи измерения качества данных.

4. **From Cleaning before ML to Cleaning for ML**

   - Традиционный подход к очистке данных не всегда эффективен.
   - Для лучшей производительности очистка данных должна интегрироваться на протяжении всего жизненного цикла модели.

Из рассмотренных исследований только у одного основным фокусом является профилирование данных, однако даже в нём функциональность инструментов рассматривается на основе статистического анализа. В остальных работах про профилирование либо ничего не говорится, либо говорится совсем мало. Поэтому, для выявления статуса наукоёмкого профилирования у приложений по работе с данными, необходимо было провести собственный обзор и сравнение.

# 3 Выбор подходящих инструментов

## 3.1 Выбор сравниваемых примитивов

Перед тем как искать подходящие инструменты, необходимо было выбрать функционал, реализация которого будет в последующем сравниваться в инструментах. Так как данная работа проводилась в рамках проекта Desbordante, было решено использовать для сравнения примитивы, поиск[2] которых реализован в данном инструменте. В результате был получен следующий список:

1. **Функциональная зависимость (FD)**

   *Определение*: Функциональная зависимость $\alpha \to \beta$ говорит о том, что для всех пар записей с одинаковыми значениями атрибутов $\alpha$, значения атрибутов $\beta$ также должны быть одинаковыми. Таким образом, значения $\alpha$ функционально определяют значения $\beta$.

2. **Условная функциональная зависимость (CFD)**

   *Определение*: CFD — это расширение FD, которое включает в себя условие, ограничивающее область действия зависимости. Например, CFD $\alpha \to \beta$ при условии $\gamma$ говорит о том, что если значения $\alpha$ и $\gamma$ одинаковые, то значения $\beta$ также одинаковы.

3. **Зависимость включения (IND)**

   *Определение*: IND между $R_i$ и $R_j$ говорит о том, что все значения в наборе атрибутов $\alpha$ присутствуют и в наборе атрибутов $\beta$, то есть $R_i[\alpha] \subseteq R_j[\beta]$. Поиск IND отсылает к определению внешних ключей.

4. **Уникальная комбинация столбцов (UCC)**

   *Определение*: UCC — это набор атрибутов $\alpha \subseteq R$, проекция которых на $r$ не содержит дубликатов. Другими словами, UCC яв-

[2]Примитивов, которые поддерживаются Desbordante, на самом деле больше, однако в ходе данной работы было решено не учитывать те из них, для которых реализована только проверка.

ляется (возможно составным) потенциальным ключом, который функционально определяет $R$.

5. **Зависимость порядка (OD)**:

   *Определение*: Зависимость порядка говорит о том, что порядок значений в одном наборе атрибутов должен соответствовать порядку значений в другом наборе атрибутов. Например, если $A$ и $B$ — два атрибута, то OD может утверждать, что если $A_i < A_j$, то $B_i < B_j$.

6. **Нечёткие алгебраические ограничения (FAC)**: *Определение*: Нечёткие алгебраические ограничения — это ограничения, которые используют алгебраические операторы для определения соответствия между значениями атрибутов. Например, если $A$ и $B$ — два атрибута, то FAC между ними утверждает, что $A$ и $B$ удовлетворяют некоторому алгебраическому выражению, такому что $A \approx B$ с заданной степенью сходства.

7. **Дифференциальная зависимость (DD)**:

   *Определение*: Дифференциальная зависимость утверждают, что разность значений в одном наборе атрибутов зависит от разности значений в другом наборе атрибутов. Например, если $A$ и $B$ — два атрибута, то DD говорит о том, что $B_i - B_j$ зависит от $A_i - A_j$ и не превышает заданного порога.

8. **Ассоциативные правила (AR)**

   *Определение*: Ассоциативные правила — это правила, которые описывают вероятность совместного появления элементов в транзакции. Например, правило $A \to B$ утверждает, что если элемент $A$ присутствует в транзакции, то с некоторой вероятностью в этой же транзакции будет присутствовать элемент $B$. Данная вероятность определяется при помощи двух метрик: поддержки и достоверности.

Часть данных примитивов рассматривалась в разобранных выше исследованиях, однако, как и было сказано выше, достаточного внимания уделено не было.

## 3.2 Выбор инструментов

Посредством изучения исследований со схожей тематикой, просмотра различных подборок и поиска в интернете было найдено 54 инструмента, которые потенциально были способны на проведение профилирования более сложного чем статистическое.

Далее были разобраны документации и демонстрации выбранных приложений. Основная направленность многих кандидатов заключалась в очистке данных, их сравнении или хранении, однако, в ходе данной работы подобный функционал не учитывался, поэтому инструменты, которые не поддерживали возможности профилирования либо оказывались пригодны исключительно для стандартного статистического анализа, были исключены из дальнейшего рассмотрения.

Необходимо сказать, что от инструментов требовалась встроенная реализация поиска выбранных примитивов. Некоторые из изначально выбранных приложений поддерживают подключение сторонних модулей и при помощи, например, библиотеки pyFPGrowth [7] могут реализовать поиск ассоциативных правил, а инструменты на подобии Apache Griffin [18], позволяющие осуществлять гибкую настройку кодовой базы, могут в теории использоваться для реализации поиска различных примитивов используя дополнительно написанный код. Однако, поскольку направленность данной работы была в обзоре встроенных реализаций, данные инструменты не были включены в итоговый список.

Также было решено не включать JuliusAI [2] в финальный список из-за отсутствующей документации и невозможности провести нормальное тестирование (количество бесплатных запросов ограничено).

По итогу осталось 12 инструментов (включая Desbordante), для которых необходимо провести сравнение.

# 4 Сравнение

Данная глава посвящена сравнению выбранных приложений. Сначала будет приведена таблица с результатами сравнения реализаций выбранных примитивов (Таблица 2), далее будет дана уточняющая информация по каждому инструменту и показана небольшая таблица, содержащая дополнительную информацию (Таблица 3). В конце будет сделан вывод, основанный на проведённом сравнении.

## 4.1 Сравнение инструментов

Таблица 2: Сравнение инструментов (примитивы)

| Инструмент | FD | CFD | IND | UCC | OD | FAC | DD | AR |
|---|---|---|---|---|---|---|---|---|
| SAP IS | + | - | - | ч | - | - | - | - |
| Informatica DQ | + | - | + | + | - | - | - | - |
| IBM InfoSphere IS | - | - | + | + | - | - | - | + |
| Experian Pandora | + | - | - | + | - | - | - | - |
| DataCleaner | - | - | ч | ч | - | - | - | - |
| Ataccama ONE | + | - | + | + | - | - | - | - |
| Oracle WB | ч | - | - | - | - | - | - | - |
| Talend | + | - | + | + | - | - | - | - |
| HoloClean | + | + | - | - | - | - | - | - |
| Uni-Detect | + | - | - | - | - | - | - | - |
| Metanome | + | + | + | + | + | - | - | - |
| Desbordante | + | + | + | + | + | + | + | + |

**SAP Information Steward [15].** Мощный инструмент для управления качеством данных, который помогает компаниям проводить контроль, улучшение и анализ имеющихся у них данных. Данная платформа позволяет устанавливать метрики и правила для оценки качества данных, отслеживать происхождение данных, а также имеет возможность создать согласованный глоссарий терминов внутри компании для упрощения взаимодействия между отделами.

Модуль профилирования данных имеет возможность проведения Dependency profile task, данная задача позволяет определить функциональные зависимости вида n:1 для заданных пользователем колонок.

Также существует Uniqueness profile task, определяющая количество неуникальных данных в таблице для заданного набора колонок, что частично можно отнести к определению уникальной комбинации столбцов.

**Informatica Data Quality [12].** Гибкий инструмент, позволяющий организациям повышать качество и проверять соответствие между данными при помощи встроенных модулей очистки и стандартизации. Платформа предлагает библиотеку готовых бизнес-правил, которые можно сразу начать применять.

В контексте данной работы нас интересуют следующие возможности: Functional Dependency Discovery, Primary Key Discovery и Foreign Key Profiling. Functional Dependency Discovery автоматически находит функциональные зависимости вида n:1, в свою очередь Primary Key Discovery позволяет найти комбинацию колонок, являющуюся уникальной для заданного набора. Foreign Key Profiling показывает пользователю возможные внешние ключи, а пользователь может одобрить или отклонить данную зависимость.

**IBM InfoSphere Information Server [11].** Комплексный инструмент для управления качеством данных, включает в себя большое количество модулей, позволяющих решать различные задачи. Ориентирован на работу с большими объёмами данных и поддерживает интеграцию с другими решениями от IBM.

Модуль Information Analyzer позволяет запустить комплексную задачу Discover and analyze primary-foreign keys, которая выявляет зависимости между колонками разных таблиц. Данный процесс позволяет определить внешние ключи, а также выявить лучший уникальный ключ (возможно составной).

Благодаря интеграции с IBM Watson имеется возможность запустить алгоритм поиска ассоциативных правил.

**Experian Pandora [8].** Платформа для управления качеством данных, очистки и обогащения. Ориентирован на автоматизацию процессов, относящихся к качеству данных. Имеет интуитивно понятный интерфейс с большим количеством настроек, что позволяет пользовате-

лям создавать собственные правила для проверки.

Модуль Dependency & Key Analysis позволяет найти функциональные зависимости вида n:1, а также выявить потенциальный составной ключ для каждой таблицы.

**DataCleaner [3].** Проект с открытым исходным кодом, направленный на удобную организацию, анализ и очищение данных. Основной целью является упрощение процесса работы с данными, делая его доступным для большего количества пользователей.

Помимо статистического анализа предоставляет только возможность проверки правильности уникального ключа и зависимости внешних ключей. Данный функционал можно частично отнести к UCC и IND.

**Ataccama ONE [1].** Платформа, которая предоставляет комплексные решения для профилирования, очистки, обогащения и стандартизации данных. Использует машинное обучение для автоматизации процессов проверки и очистки данных, также имеет встроенные инструменты для визуализации и проведения аналитики, включая отслеживания метрик качества данных в реальном времени.

Data Profiling модуль данного приложения позволяет задать колонки и проверить для них функциональные зависимости вида n:1, проверить наличие зависимости между колонками разных таблиц (проверка внешних ключей), а также провести анализ колонок на определение уникального ключа (возможно составного).

**Oracle Warehouse Builder [14].** Инструмент для создания и управления хранилищами данных, а также обеспечивающий проверки качества данных. Oracle прекратили активную поддержку данного продукта, перенеся основной функционал в Oracle Data Integrator, однако OWB всё ещё применяется в различных ситуациях. Направленностью приложения является извлечение и трансформация данных, но существует и модуль для профилирования данных.

OWB позволяет запустить Functional Dependency Analysis для заданной пары колонок и проверить функциональную зависимость вида 1:1, что можно частично отнести к определению FD, Unique Key Analysis проверяет, является ли указанный атрибут уникальным клю-

чом, однако, проверяемый ключ может состоять только из одной колонки, так что данный функционал не относится к UCC даже частично.

**Talend [17].** Платформа предоставляющая инструменты для управления большими данными, их обработки и очистки. Раньше была разделена на коммерческое решение и часть с открытым исходным кодом, которая называлась Talend Open Studio, однако начиная с 31 января 2024 года поддержка бесплатной версии была прекращена.

При помощи Potentional Primary Keys можно найти потенциальные уникальные составные ключи, Functional Dependency Analysis позволяет проверить, удерживается ли функциональная зависимость вида n:1 для заданных пользователем колонок. Помимо этого существует возможность проверки зависимости внешнего ключа, что можно отнести к поиску IND.

**HoloClean [10].** Интересное решение, которое направлено на восстановление и улучшение данных при помощи машинного обучения и минимального участия пользователя. Данный инструмент находит функциональные зависимости, условные функциональные зависимости и на основе вероятностных ограничений, создаваемых автоматически, исправляет аномалии, пропущенные значения и неточности.

**Uni-Detect [19].** Инструмент, который предназначен для автоматического поиска ошибок. При помощи поиска функциональных зависимостей и «what-if» анализа Uni-Detect определяет потенциальные неточности в данных, о которых сообщает пользователю.

**Metanome [13].** Metanome представляет из себя инструмент с открытым исходным кодом, содержащий большое количество алгоритмов для поиска примитивов, направленных на улучшение данных.

Среди реализованных алгоритмов есть алгоритмы поиска FD, CFD, IND, UCC, OD, а также некоторых других примитивов, не рассматривающихся в рамках данной работы.

**Desbordante [4].** Проект, изначально вдохновленный Metanome, однако имеющий собственное видение данной сферы. Позволяет пользователям запускать алгоритмы поиска отмеченных примитивов, а также различных сценариев по очистке и улучшению данных через различ-

ные интерфейсы (интерфейс командной строки, питон-привязки, веб-приложение).

Таблица 3: Сравнение инструментов (дополнительно)

| Инструмент | Происхождение | Интерфейс | Год |
|---|---|---|---|
| SAP IS | Industrial | WEB | 2011 |
| Informatica DQ | Industrial | WEB; Desktop | 2006 |
| IBM InfoSphere IS | Industrial | WEB | 2006 |
| Experian Pandora | Industrial | Desktop | 2015 |
| DataCleaner | Open-source | Desktop; API | 2009 |
| Ataccama ONE | Industrial | WEB | 2018 |
| Oracle WB | Industrial | Desktop | 2000 |
| Talend | Industrial | WEB; Desktop | 2005 |
| HoloClean | Open-source | API | 2018 |
| Uni-Detect | Статья | CLI | 2019 |
| Metanome | Open-source | WEB | 2015 |
| Desbordante | Open-source | WEB; CLI; API | 2021 |

Перед подведением итогов следует уточнить, что помимо профилирования данных к области Data Quality относятся и другие задачи, поэтому будет справедливо в краткой форме изложить их поддержку в рассматриваемых приложениях.

Основные задачи Data Quality были взяты из первого исследования [5], рассматриваемого ранее. К ним относится:

- Профилирование данных (далее DP);
- Измерение качества данных (далее DQ Measur.);
- Очистка данных (далее DC);
- Мониторинг качества данных (далее DQ Monit.);
- Управление качеством данных (далее DQ Manag.).

Поддерживаемый инструментами функционал указан в соответствующей таблице (Таблица 4).

Таблица 4: Сравнение инструментов (data quality tasks)

| Инструмент | DP | DQ Measur. | DC | DQ Monit. | DQ Manag. |
|---|---|---|---|---|---|
| SAP IS | + | + | - | - | + |
| Informatica DQ | + | + | + | - | + |
| IBM InfoSphere IS | + | + | - | + | + |
| Experian Pandora | + | + | + | - | - |
| DataCleaner | + | - | + | - | - |
| Ataccama ONE | + | + | + | + | + |
| Oracle WB | + | + | - | - | - |
| Talend | + | - | + | - | + |
| HoloClean | + | - | + | - | - |
| Uni-Detect | + | - | + | - | - |
| Metanome | + | - | - | - | - |
| Desbordante | + | - | - | - | - |

## 4.2 Результаты

Как видно из таблицы (Таблица 2), приложениями в основном поддерживается только определение функциональных зависимостей, определение зависимостей включения и определение уникальных комбинаций столбцов, причём инструментов, которые могут автоматически находить эти примитивы без ручного ввода пользователем, ещё меньше.

Также, если проанализировать колонку «Происхождение» из таблицы (Таблица 3), можно сделать вывод, что основная часть инструментов (7 из 11) представляет собой коммерческое решение, что сказывается на доступности рассматриваемого вида профилирования.

Данное сравнение показывает, что доступных инструментов, подходящих для наукоёмкого профилирования мало, а тех, которые реализуют поиск примитивов более сложных чем функциональные зависимости, единицы.

# Заключение

В рамках данной работы были достигнуты следующие результаты:

1. Проведён обзор существующих исследований.
2. Отобраны подходящие для сравнения инструменты.
3. Проведён обзор отобранных инструментов.
4. Проведено сравнение отобранных инструментов.
5. Сделаны выводы о статусе наукоёмкого профилирования среди отобранных инструментов.

# Список литературы


[1] Ataccama ONE. — 2024. — URL: https://www.ataccama.com/platform.

[2] Caesar Labs, Inc. Julius: Large Language Model. — https://julius.ai/. — 2024.

[3] DataCleaner. — 2024. — URL: https://datacleaner.github.io/.

[4] Chernishev George, Polyntsov Michael, Chizhov Anton et al. Desbordante: from benchmarking suite to high-performance science-intensive data profiler (preprint). — 2023. — 2301.05965.

[5] Ehrlingerand Lisa, Wöß Wolfram. A Survey of Data Quality Measurement and Monitoring Tools. — 2022. — URL: https://doi.org/10.3389/fdata.2022.850611.

[6] Evaluation of Freely Available Data Profiling Tools for Health Data Research Application: A Functional Evaluation Review / B Gordon, C Fennessy, S Varma et al. // BMJ Open. — 2022. — Vol. 12, no. e054186. — URL: https://doi.org/10.1136/bmjopen-2021-054186.

[7] Evan Dempsey. FP-Growth: A Python Implementation of the Frequent Pattern Growth Algorithm. — https://github.com/evandempsey/fp-growth. — 2024.

[8] Experian Pandora. — 2024. — URL: https://www.experian.com/data-quality/experian-pandora.

[9] From Cleaning before ML to Cleaning for ML / Felix Neutatz, Binger Chen, Ziawasch Abedjan, Eugene Wu // IEEE Data Eng. Bull. — 2021. — Vol. 44. — P. 24–41. — URL: https://api.semanticscholar.org/CorpusID:237542697.

[10] Rekatsinas Theodoros, Chu Xu, Ilyas Ihab F., Ré Christopher. HoloClean: Holistic Data Repairs with Probabilistic Inference. — 2017. — 1702.00820.

[11] IBM InfoSphere Information Server. — 2024. — URL: https://www.ibm.com/information-server.

[12] Informatica Data Quality. — 2024. — URL: https://www.informatica.com/products/data-quality.html.

[13] Metanome. — 2024. — URL: https://hpi.de/naumann/projects/data-profiling-and-analytics/metanome-data-profiling.html.

[14] Oracle Warehouse Builder. — 2024. — URL: https://www.oracle.com/application-development/technologies/warehouse/warehouse-builder.html.

[15] SAP Information Steward. — 2024. — URL: https://www.sap.com/products/technology-platform/data-profiling-steward.html.

[16] Zhou Yuhan, Tu Fengjiao, Sha Kewei et al. A Survey on Data Quality Dimensions and Tools for Machine Learning. — 2024. — 2406.19614.

[17] Talend. — 2024. — URL: https://www.talend.com.

[18] The Apache Software Foundation. Apache Griffin: A Data Quality Solution For Hadoop and Spark. — https://griffin.apache.org/. — 2024.

[19] Wang Pei, He Yeye. Uni-Detect: A Unified Approach to Automated Error Detection in Tables // Proceedings of the 2019 International Conference on Management of Data. — SIGMOD '19. — New York, NY, USA : Association for Computing Machinery, 2019. — P. 811–828. — URL: https://doi.org/10.1145/3299869.3319855.

[20] YData. ydata-profiling: Data quality profiling & exploratory data analysis for Pandas and Spark DataFrames. — https://github.com/ydataai/ydata-profiling. — 2024.