**Лабораторная работа №8**

**Невейков Андрей, 2022**

**All scripts here:**

## **https://github.com/AndreyNeveikov/DataMola\_Labs/tree/main/Modul\_2/lab8/package\_for---t\_dw\_fct\_orders**

## Lab8: 2.1. Task 01: Extraction Description

**Full Extraction**

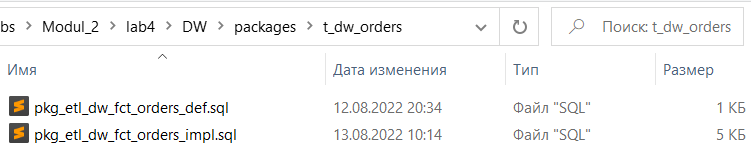
Суть метода: данные полностью извлекаются из исходной системы. Поскольку это извлечение отражает все данные, доступные в настоящее время в исходной системе, нет необходимости отслеживать изменения в источнике данных с момента последнего успешного извлечения.   
 Этот способ в бизнесе может использоваться при получении данных о заказе с сайта: т.к. эти таблицы хранят мало информации, а скорость выполнения заказа очень важна для привлечения и сохранения клиента выгоднее постоянно обновлять информацию о заказе. Быстрее получим информацию -> быстрее начнем выполнять заказ -> быстрее отдадим заказ клиенту -> клиент сильнее захочет купить что-нибудь еще.

**Offline Extraction**

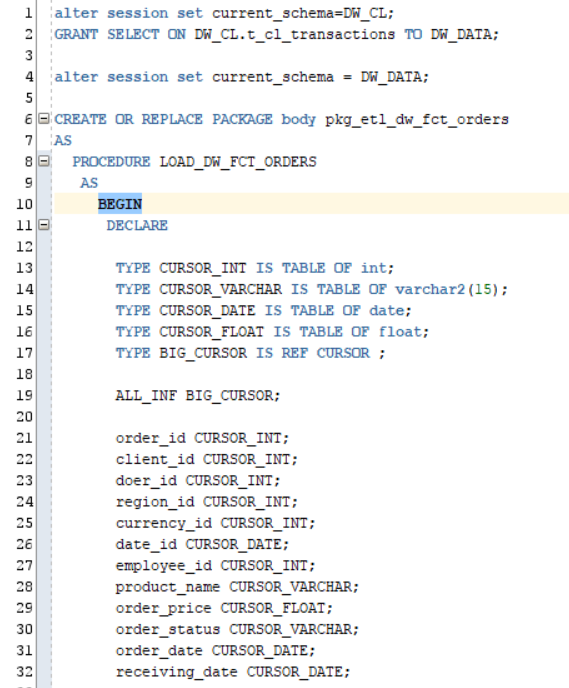
Суть метода: данные не извлекаются непосредственно из исходной системы, а с отдельно выделенных мощностей, чтобы не нагружать основную систему и не блокировать к ней доступ из-за большой загрузки. В бизнесе может использоваться при составлении финансовых отчетов для налоговой, когда необходима точность в подсчете большого объема транзакций за длительный период (например: квартал, финансовый год). Выгоднее проверить не изменились ли исторические данные, не нагружая основную сеть, т.к. подобная проверка может быть выполнена на меньших мощностях, но за большее количество времени.

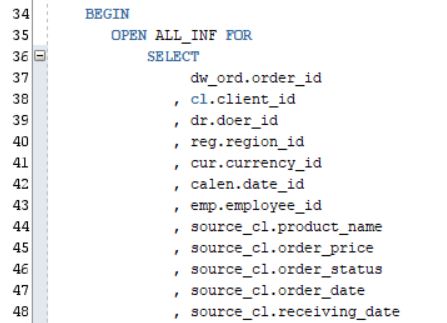
## Lab8-9 3.1. Task 02: Prepare Table of Facts to DW Layer

Пакеты для прогрузки таблицы фактов были закончены еще утром 13.08.2022 для lab4. Пакет состоит из def и impl частей. В def – объявление пакета, в impl – логика.

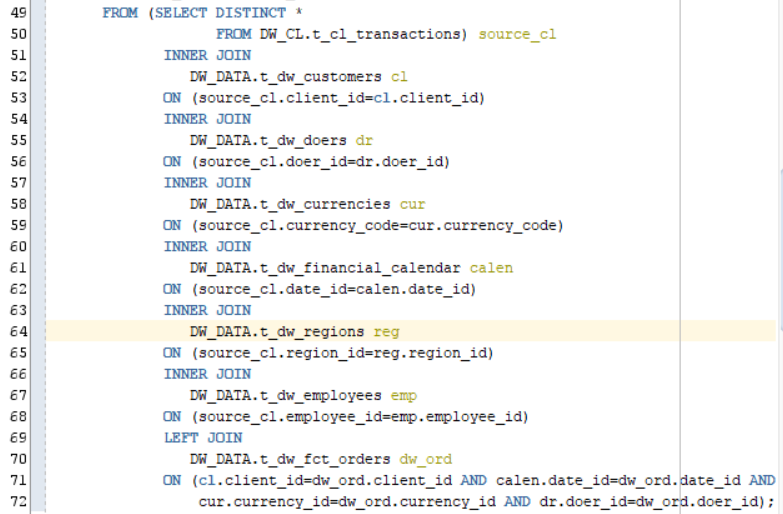


Для перемещения данных используется процедура с курсором:

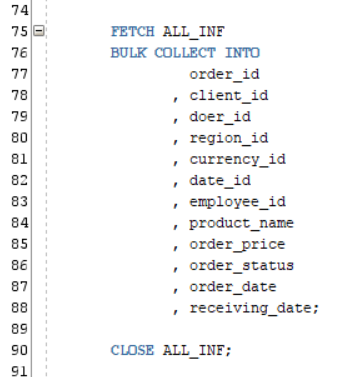




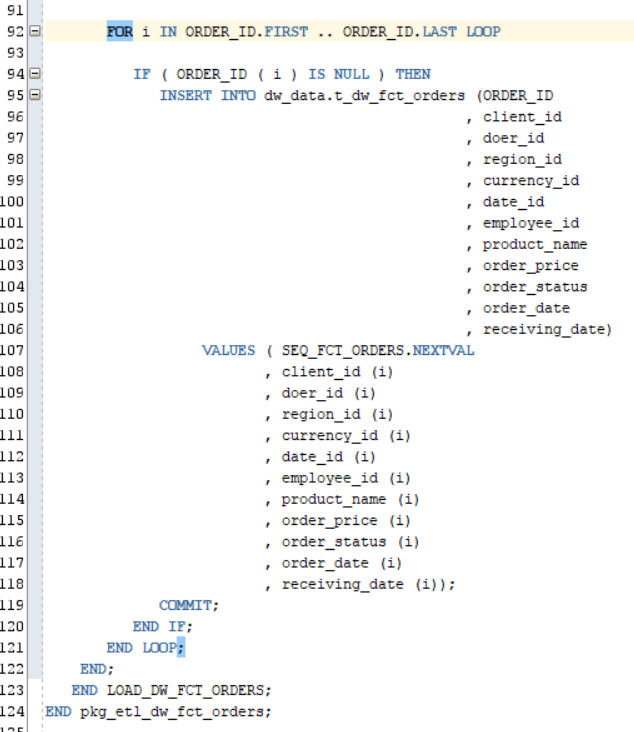
Перемещение данных происходит при помощи соединения с другими таблицами, для сохранения целостности данных т.к. на DW-уровне появляются несколько полей с индексами из других таблиц.



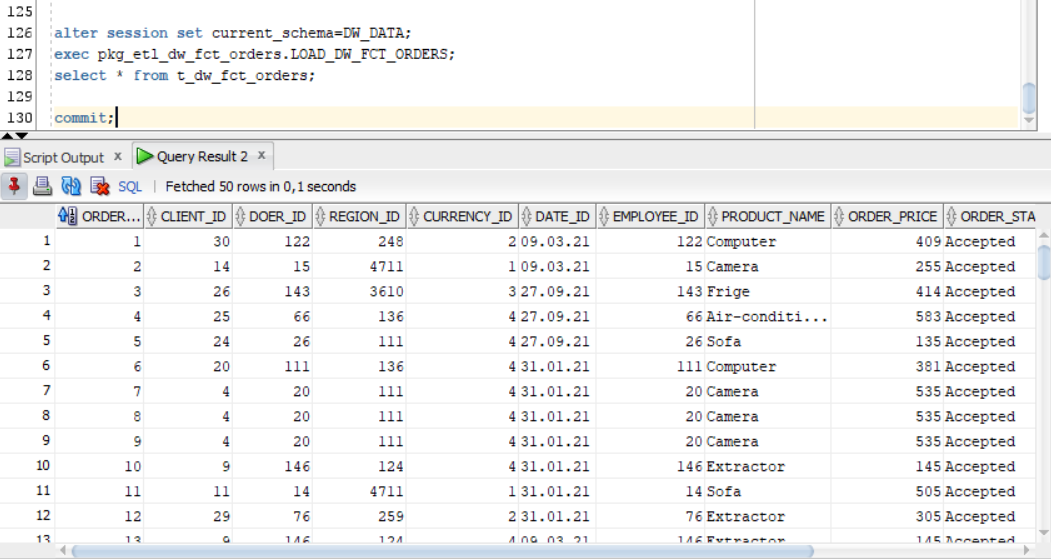
Извлекаются все данные:



Начинается цикл, который заполняет таблицу нужными значениями и присваивает собственный индекс заказа:



Результат работы пакета:



**SUMMARY**

1. **Я научился** переносить данные в факт-таблицу.
2. **Я узнал**, способ сохранения целостности данных при переносе информации в факт-таблицу (связывая ее с другими, предварительно заполненными

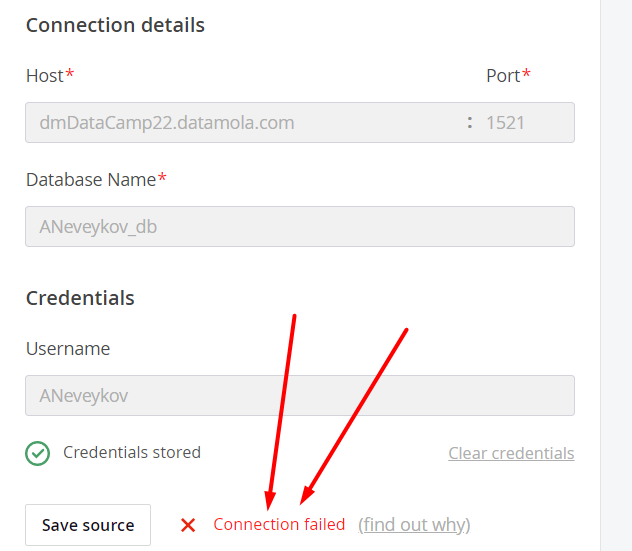
таблицами;

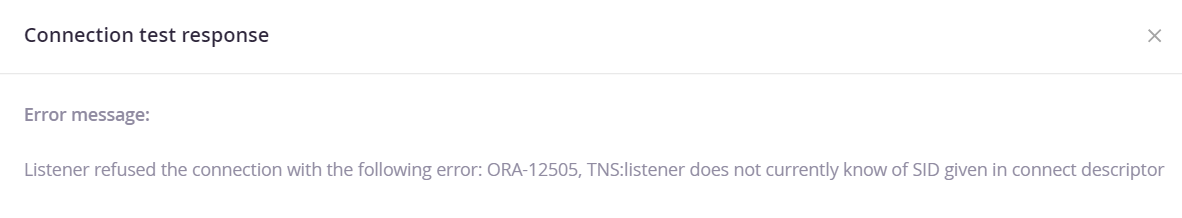
1. **Я понял**, как переносить поля из нескольких таблиц в одну одновременно.

Дополнительное задание от Кирилла:

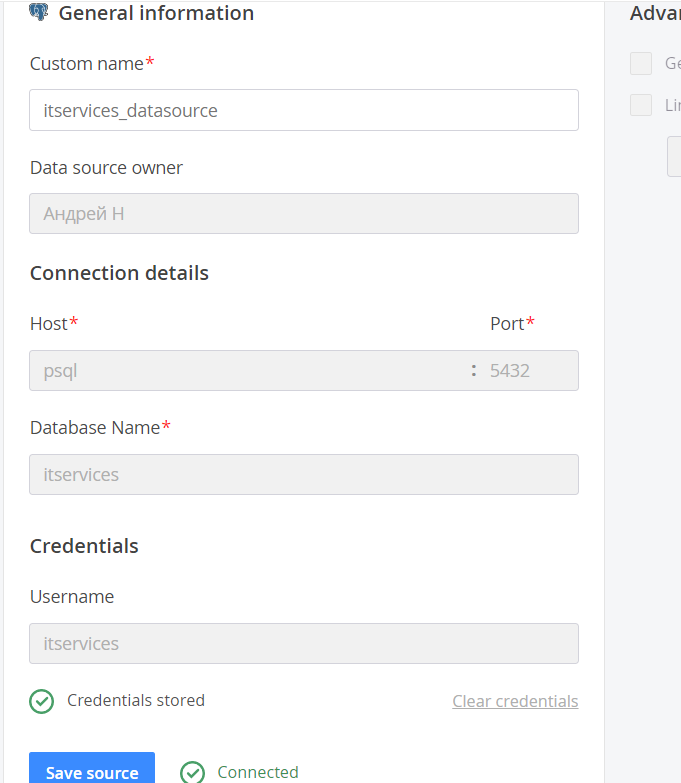
Зайти на сайт <https://app.ataccama.com/catalog> , поизучать его, понять, чем реально занимается CL-уровень.

Изначально я хотел проверить работу на данных из своей БД, но появлялась ошибка подключения:

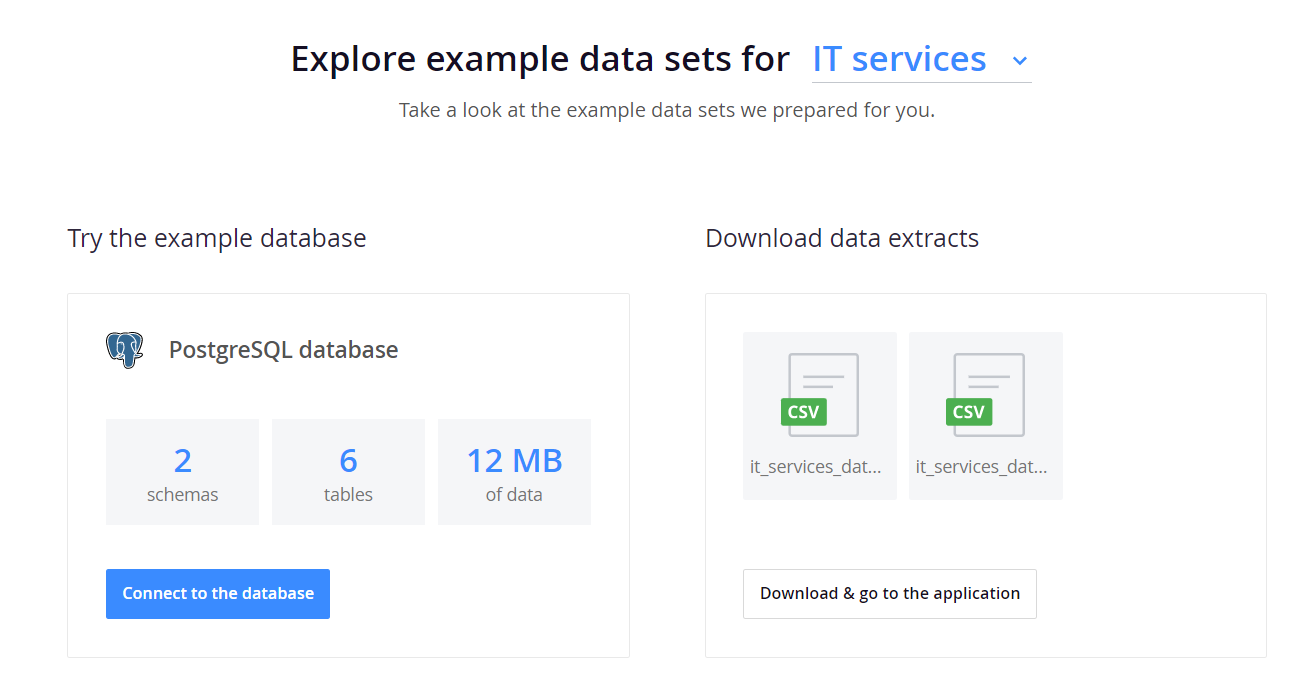




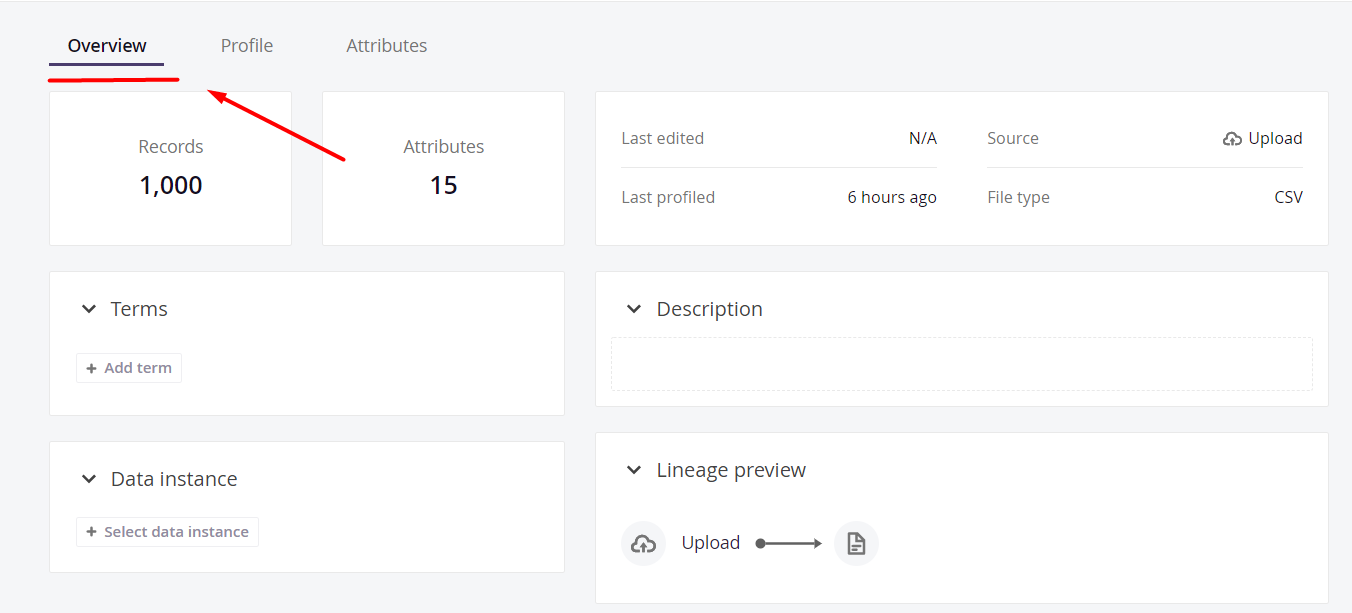
Тогда я использовал тестовый сет с их сайта:



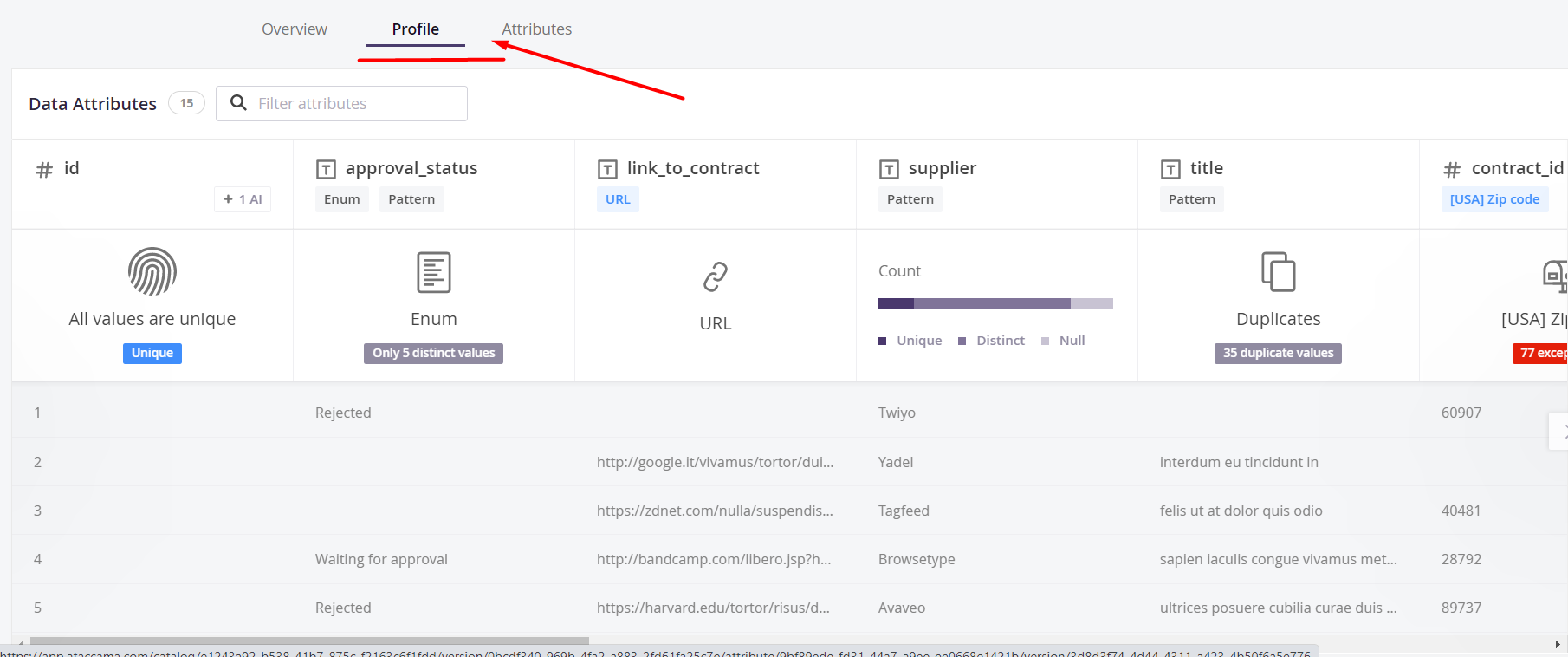
Тестовый сет состоял из 2-ух схем, 6-и таблиц на 12МВ данных:



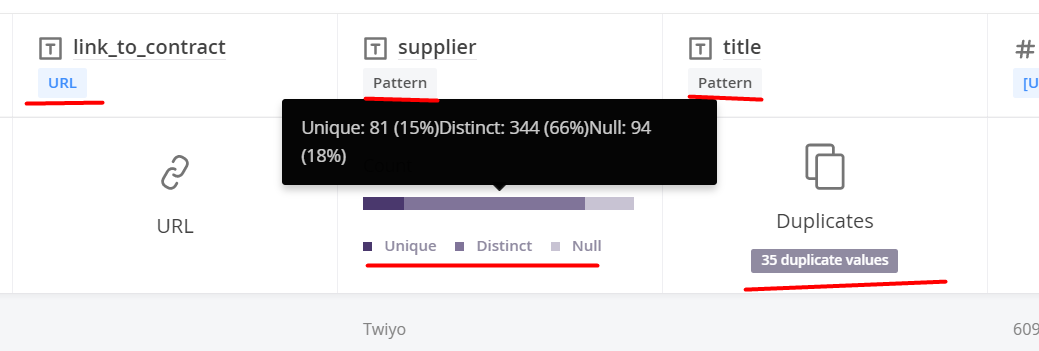
На сайте есть три основных вкладки. Overview показывает основную информацию о датасете:



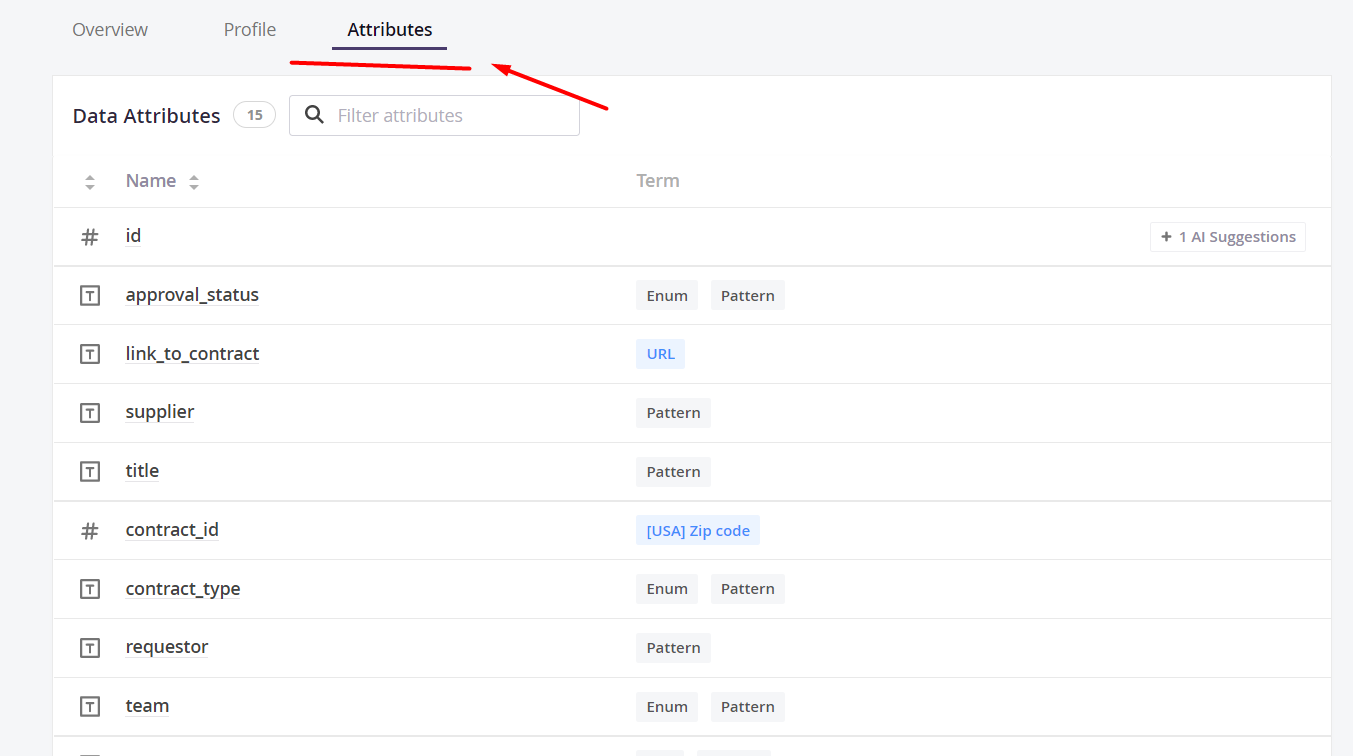
Profile дает развернутую информацию о всех колонках: определяет повторы, уникальные значения, паттерны, категории.



При нажатиях на соответствующие иконки информация становится еще подробнее:



Attributes хранит общую информацию по колонкам и предлагает подсказки, о чем могут быть эти данные:



***Таким образом,*** на CL уровне определяются значения-дубликаты, сравнивается формат каждой колонки с шаблоном, происходит анализ уникальных столбцов и поиск категориальных переменных, определяется процент null-значений.