**Anton Slizh’s**

**U2M8.LW.ETL Overview – Extraction**

**&**

**Ataccama platform overview**

*GitHub: https://github.com/drapejny/DataCamp2022*

**Task 1**

## 2.1. Task 01: Extraction Description

To choose the extraction method we should consider different factors. The main factor is the possibilities of source system to implement one or another method. Very often, there is no possibility to add additional logic to the source systems to enhance an incremental extraction of data due to the performance or the increased workload of these systems. Also, we should look at the volume and the changeability of data being used.

In our business model we decided to implement the *Full Extraction* method. The main idea of this method is to extract full data completely from the source system. There is no need to keep track all data changes at definite time period. The decision was made because the volume of data system is not significantly large and there are some difficulties in implementing the *Incremental Extraction* method in provided source system.

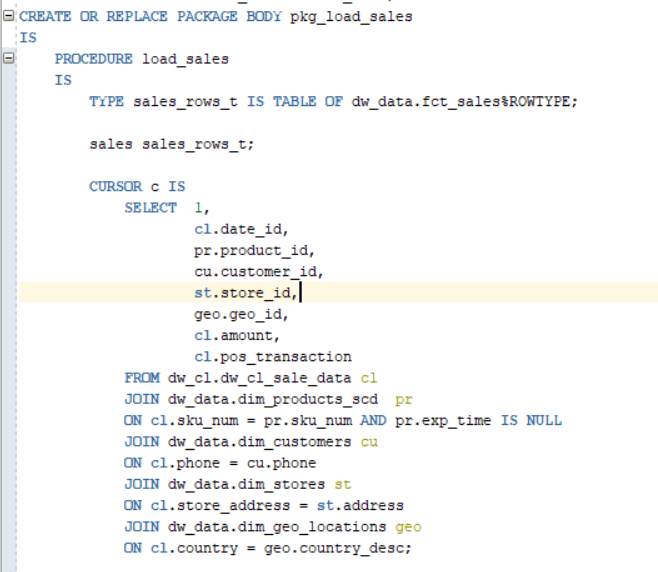
As the physical extraction method we decided to use *Offline Extraction*. Here the data is not extracted directly from the source, but instead it’s taken from another external area which keeps the copy of source. This help us avoid overloading of the main source system fetching the records from the external source instead of the actual source.

**Task 2**

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

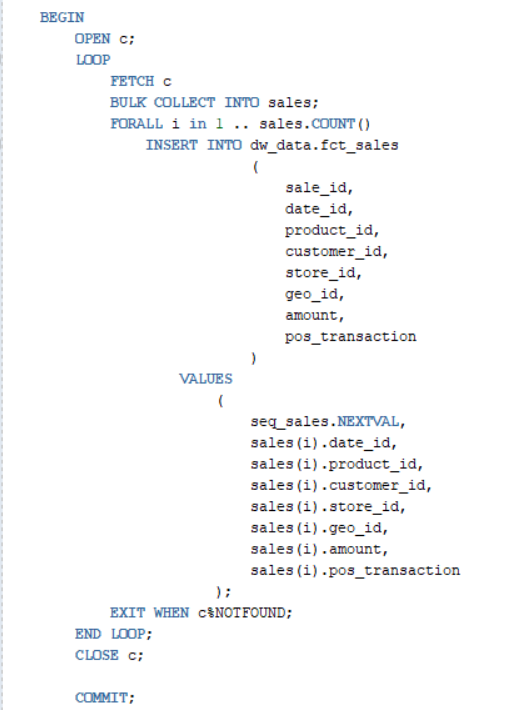
I have already created the package for load my fact table in Lab 4. Let’s look at the loading fact table *(fct\_sales)* again.

Defining procedure and variables:

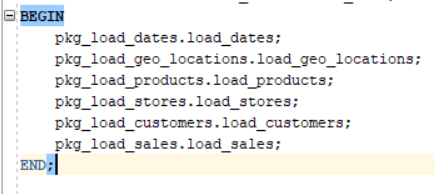


I have used the cursor to iterate throw the data. The cursor was created as select statement on the sales data from cleansing layer and joined data from dimensions (just to convert natural keys to surrogate).

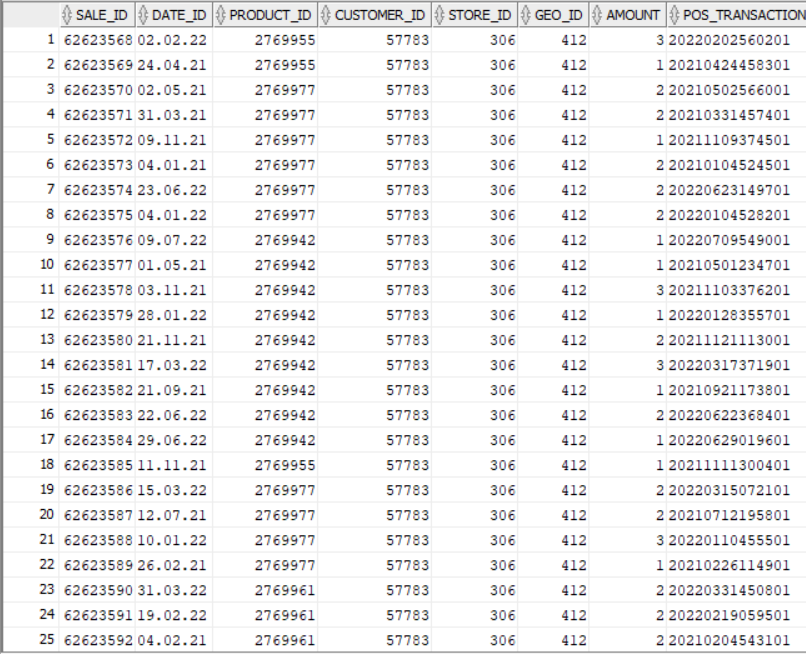
The procedure body contains bulk collecting to the sales variable and further bulk insertion into fact table.



The load sales procedure executing after updating data in the dimensions:



Result data:



**Task 3**

**Overview of Ataccama**

First of all I have registered on Ataccama (<https://app.ataccama.com/catalog>). After creating the account, I can start to analyze data.

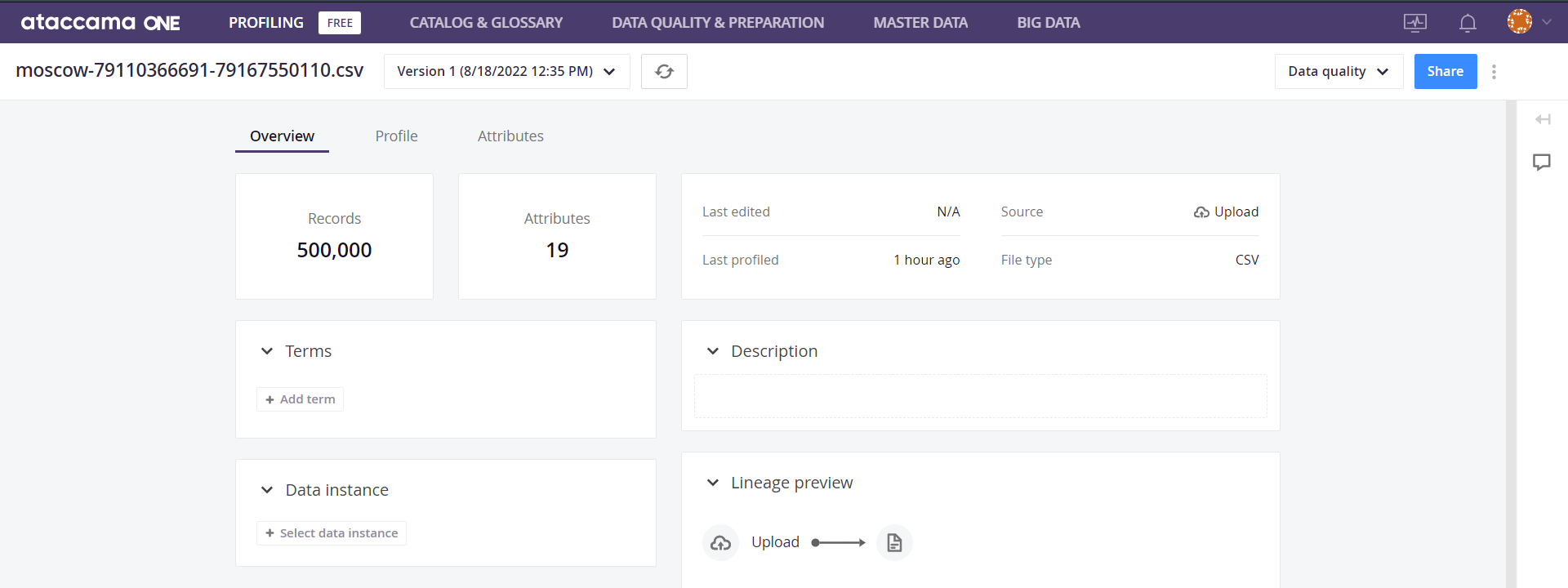
I couldn’t connect to my Oracle DB. So, I decided to manually upload the dataset to analyze.

I have the dataset of orders from one of the most popular delivery services in Russia. Few months ago, when suddenly this information became public, I downloaded it just for educational purposes. So, the moment has come.

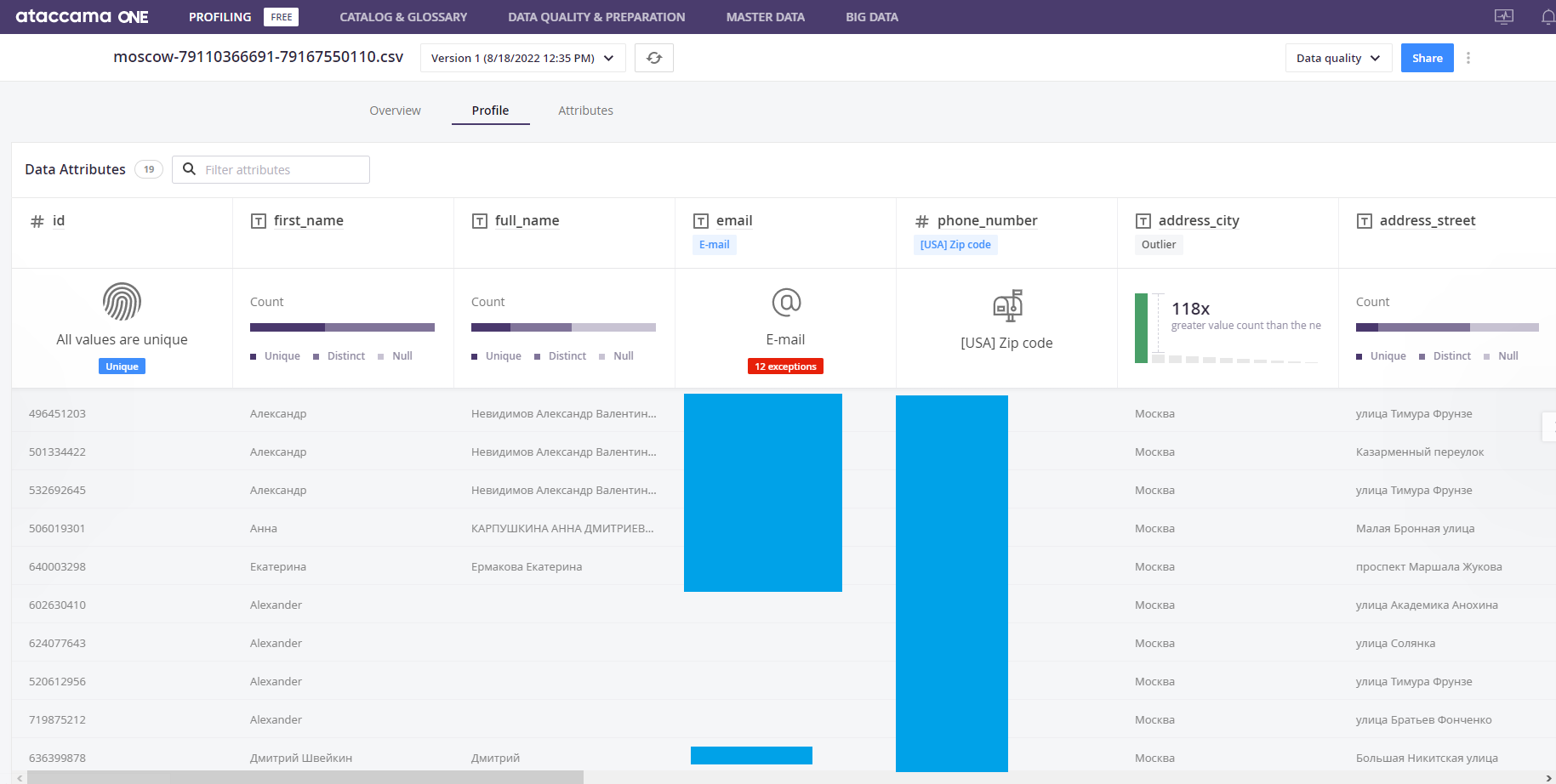
Uploading csv file with orders in Moscow region and starting the profiling process.

After profiling was completed let’s look at the prepared report about dataset.

The file contains 500.000 rows and 19 attributes.



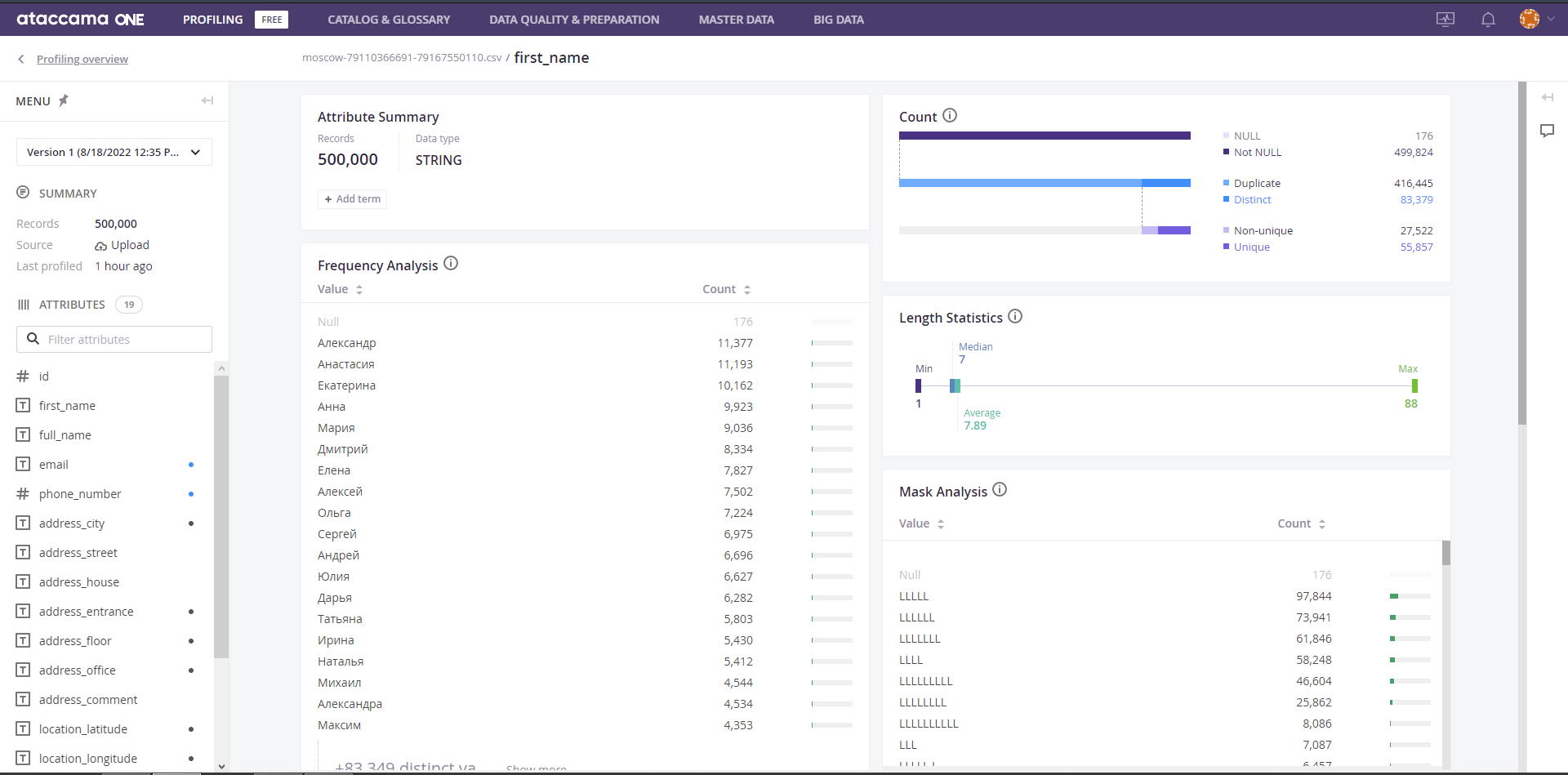
Here the overview of the dataset



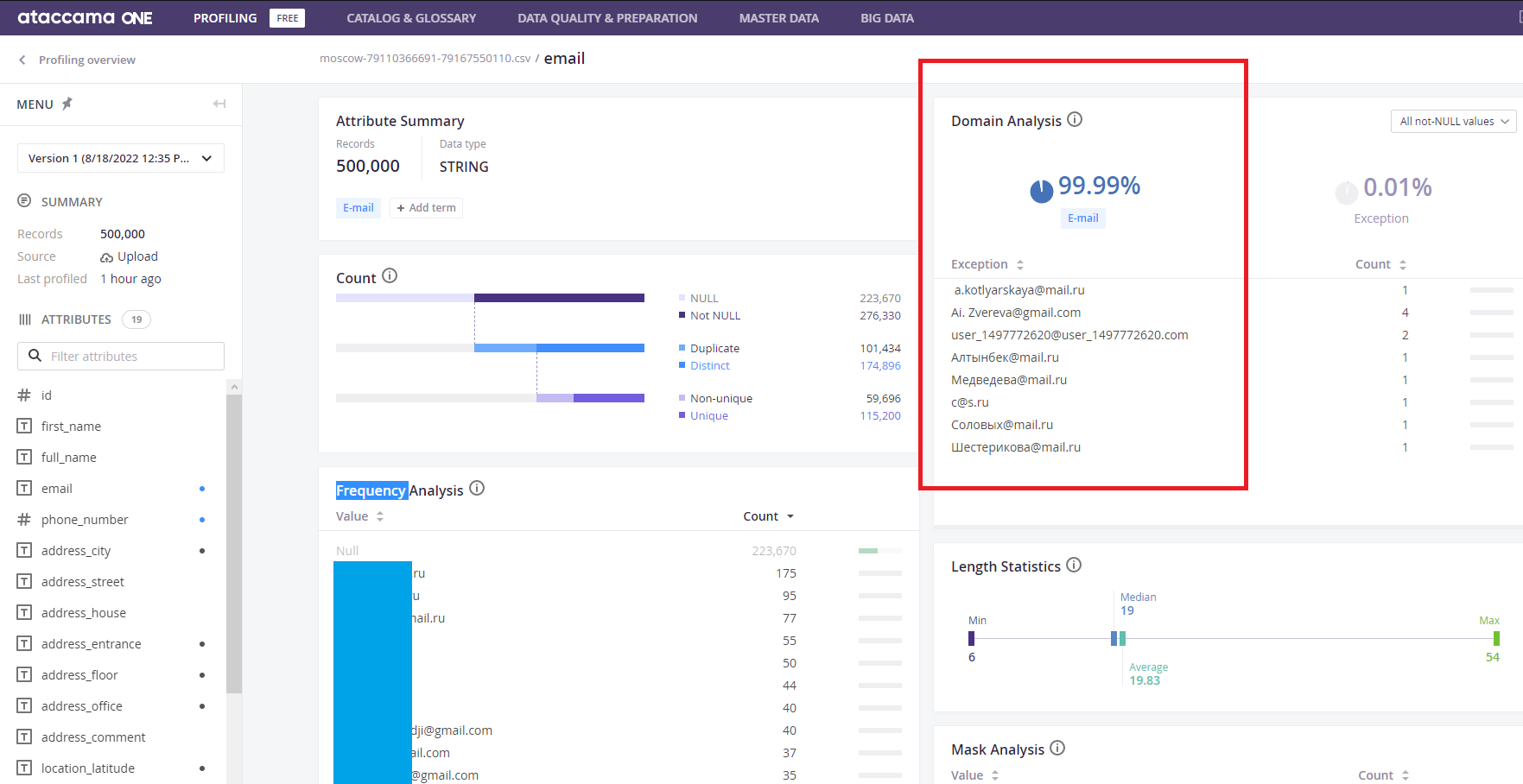
Then we can check statistics about each attribute in the dataset.

Let’s look at the *first\_name* attribute statistic.

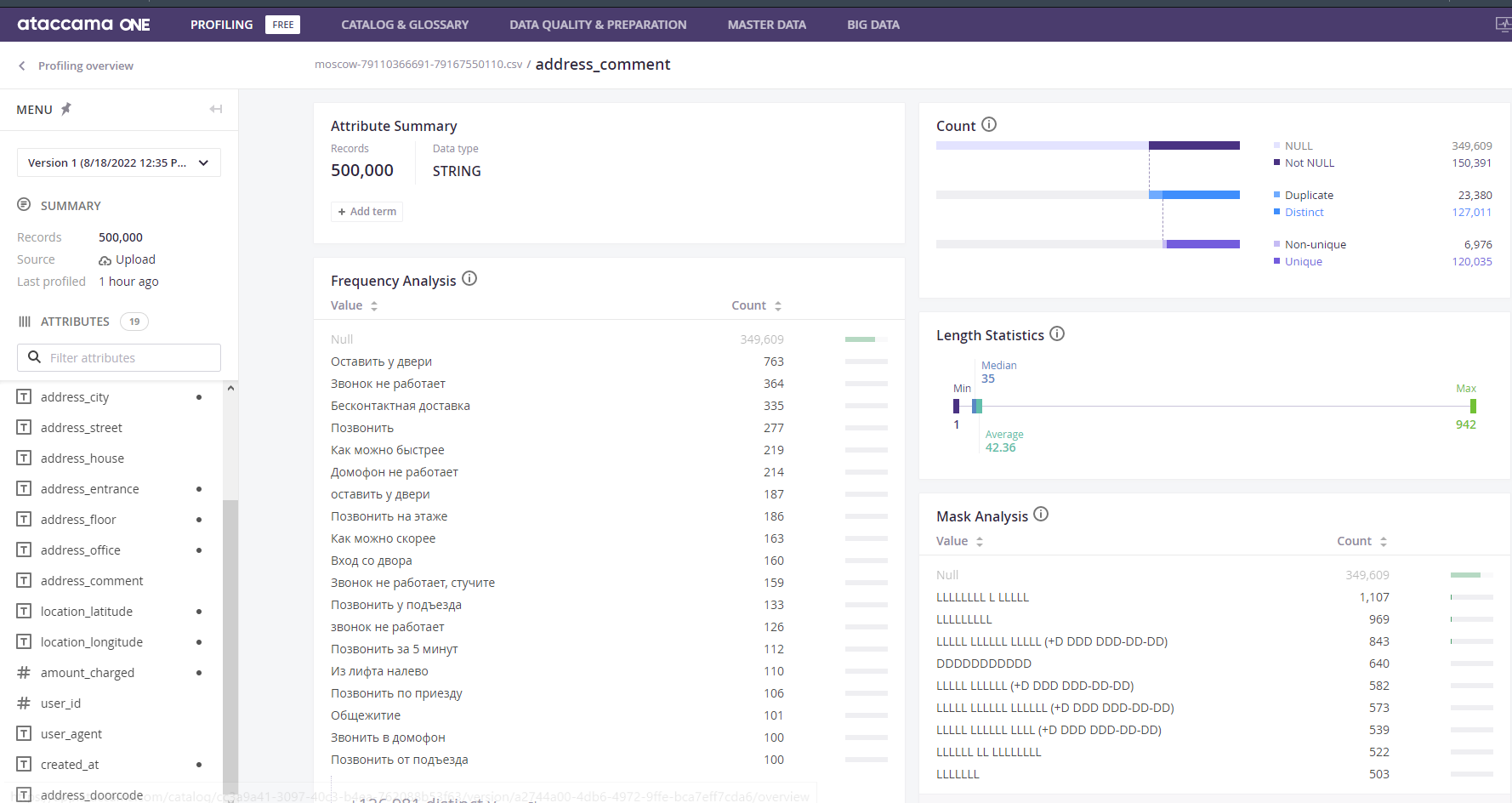
We can check the most popular first names of customers, the length and mask statistics of the attribute values.



Also let’s look at the *email* attribute. As well as the value and count statistics we can look at the exception values. In the *Domain Analysis* we can look at the exceptions and their frequency.



And let’s look at the *address\_comment* attribute. Here we can see the most popular customers comments for orders.



So, the Ataccama platform (free version) provide great tools for basic analysis of your datasets.