INFO7374 - Team 6

Alibaba Dataset

[**About Alibaba**](#_osapyzlkw3bd)

[**What are we doing for Marketa?**](#_wkzfnc6ylj6j)

[What?](#_sew7rv6ox3j6)

[Why?](#_1f66uplwev6o)

[Where?](#_cfuskcs60fvj)

[**Design Implementation**](#_dp8rg7xvt0kr)

[**Pandas/Python**](#_tz7gq35tz8yv)

[How we used Pandas?](#_lkueg8mu0531)

[What are the pros?](#_ot7257sybf8l)

[Big Spender & Many Orders](#_15xcaa12czr3)

[Faker Implementation](#_fbotdkb7go47)

[**XSV**](#_in6ayhlx2sm6)

[How we used XSV?](#_7dn9irob1tfh)

[What are the pros?](#_lkueg8mu0531)

[**Trifacta**](#_rovvhk6q5v1g)

[How we used Trifacta?](#_e9k7rq2rnh35)

[What are the pros?](#_piwhsskdg82d)

[**Salesforce einstein analytics**](#_rdcq1eknnfhk)

[How we used Einstein Analytics?](#_brh3tac724jw)

[What are the pros?](#_7xm4srjz8oti)

[Dimensions and Measures](#_degtr9vvfv9t)

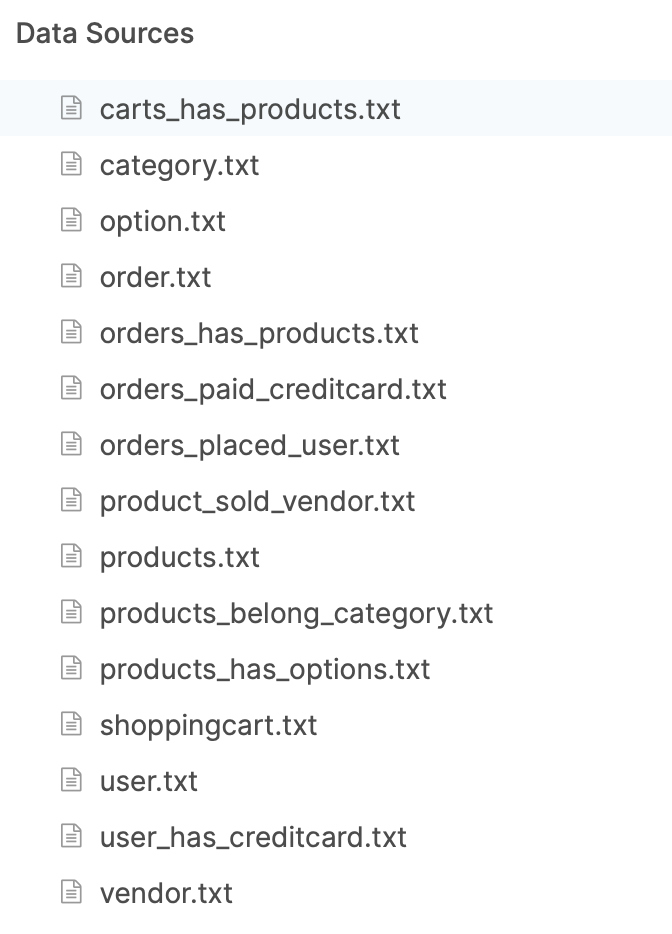
[Who would use the dashboard?](#_arxr3frzjou5)

# About Alibaba



Started in 1999, Alibaba.com is the leading platform for global wholesale trade. We serve millions of buyers and suppliers around the world.

Our dataset in Alibaba consisted of :



# What are we doing for Marketa?

## 

## What?

Illustrate the value of data driven analytics with the following themes to be included

* Pricing
* Promotion
* Search
* Recommendations

We have focussed on Sales Insights and Target Email Marketing to provide promotions & recommendations for customers of Marketa

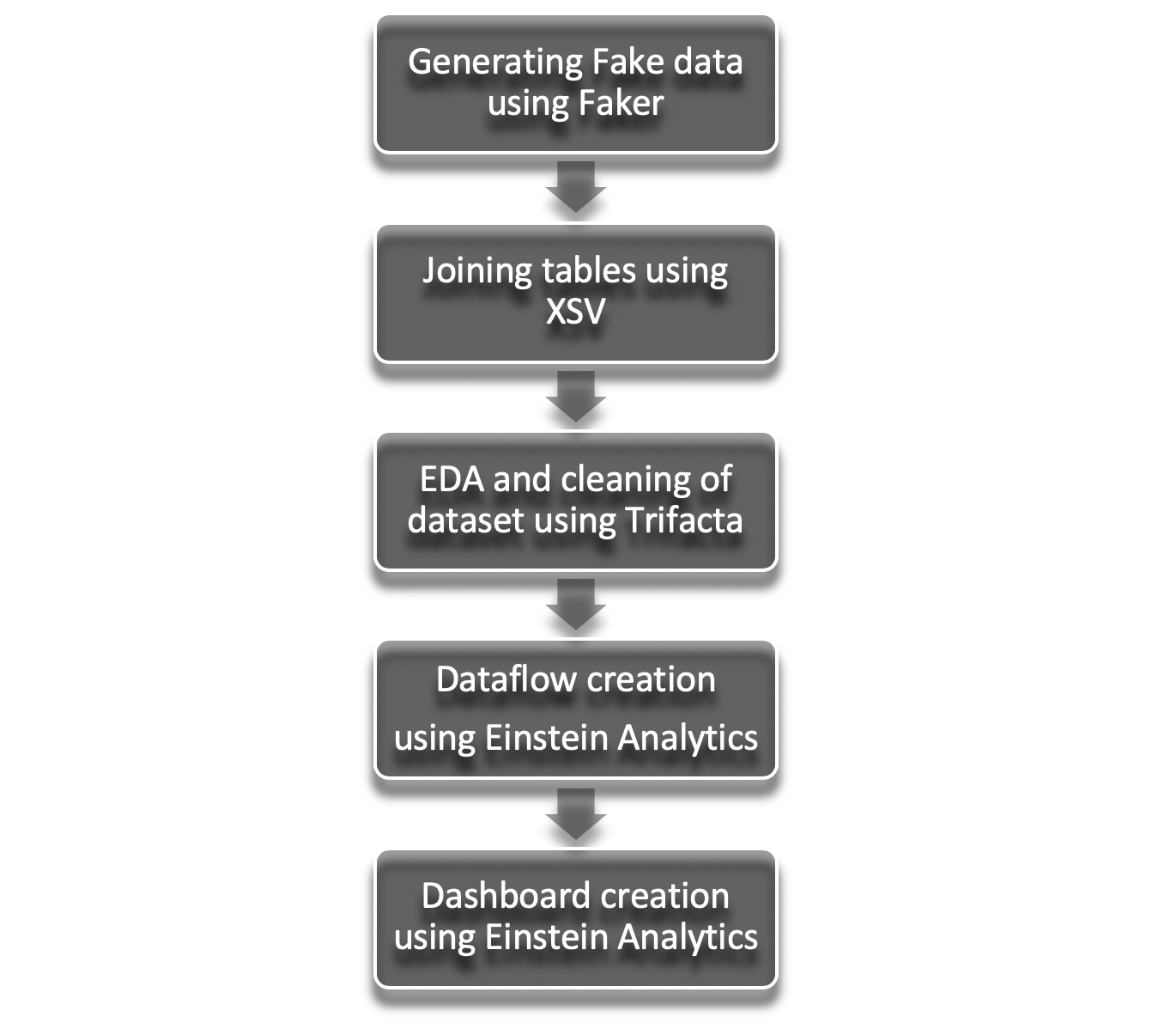
## Where?

Marketa wants you to analyze the data using tools (Pandas, xsv, Trifacta) and build a dashboard using Einstein analytics

## Why?

* To work with datasets using Pandas, xsv(xcsv) and Trifacta tools
* To be able to analyze marketing data using Salesforce Einstein analytics studio
* Derive insights from the datasets
* Crisply communicate and document your findings

# Design Implementation



# 

# 

# Pandas/Python

## How we used Pandas for our dataset?

**- Converted the .txt files to .csv via pandas**

**- Used Faker to generate data according to the format given**

## What are the pros of using Pandas?

- Pandas deals with the data processing and analysis in five steps:

* Load
* Prepare
* Manipulate
* Model
* Analyze

- Easy handling of missing data - representing it as NaNs

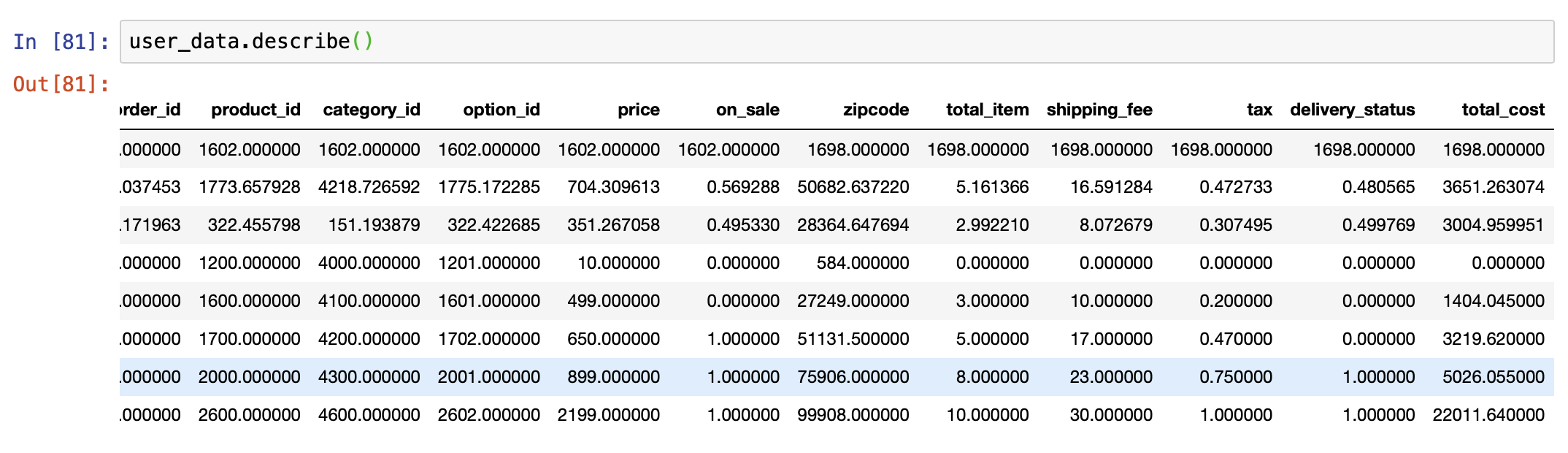
- Label-based slicing, fancy indexing, and subsetting of large data sets, as well as data set merging and joining

## Python to show the Big Spender and Many Orders as a flag

Big Spender is a customer in our dataset who is more than 75th percentile of the total\_cost

Whereas

Many Orders is a customer in our dataset who is in the 75th percentile of the total\_item

****



## 

## [Faker](https://pypi.python.org/pypi/Faker) is a Python package that generates fake data for you

The goal: given a target dataset (for example, a CSV file with multiple columns), produce a new dataset such that the anonymized dataset should have the same amount of data and maintain its analytical value



# XSV

xsv is a command line program for indexing, slicing, analyzing, splitting and joining CSV files. Commands should be simple, fast and composable:

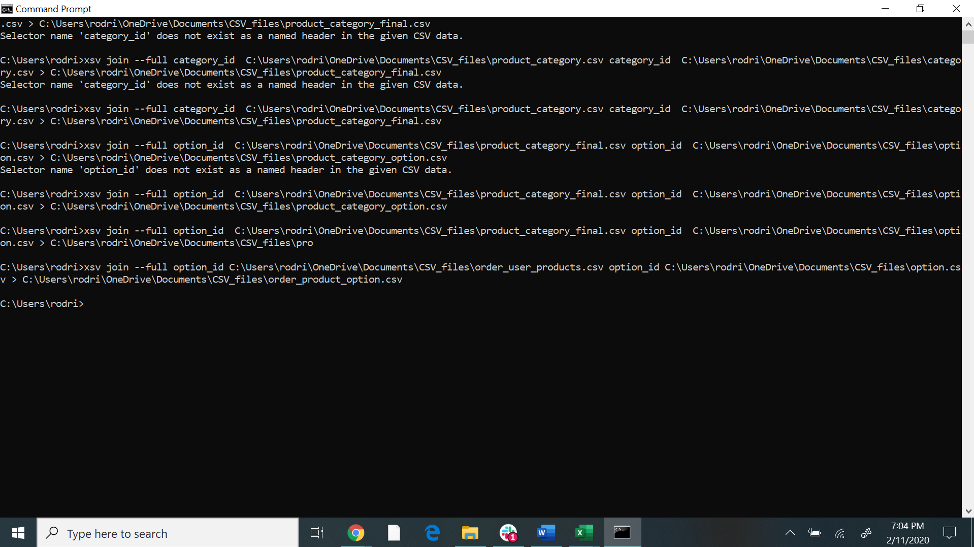
1. Simple tasks should be easy.
2. Performance trade offs should be exposed in the CLI interface.
3. Composition should not come at the expense of performance.

## How we used XSV for our datasets?

**We extensively used XSV to join our datasets , it was quicker and easier than Salesforce Einstein analytics.**

## What are the pros of XSV?

* CSV data is annoying because CSV is a terrible format- according to Burnt sushi
* If your data is gigabytes in size, then CSV is the wrong storage type





Tables used: vendor, product\_has\_options, product\_sold\_vendor, option, product\_belong\_category, category

# Trifacta

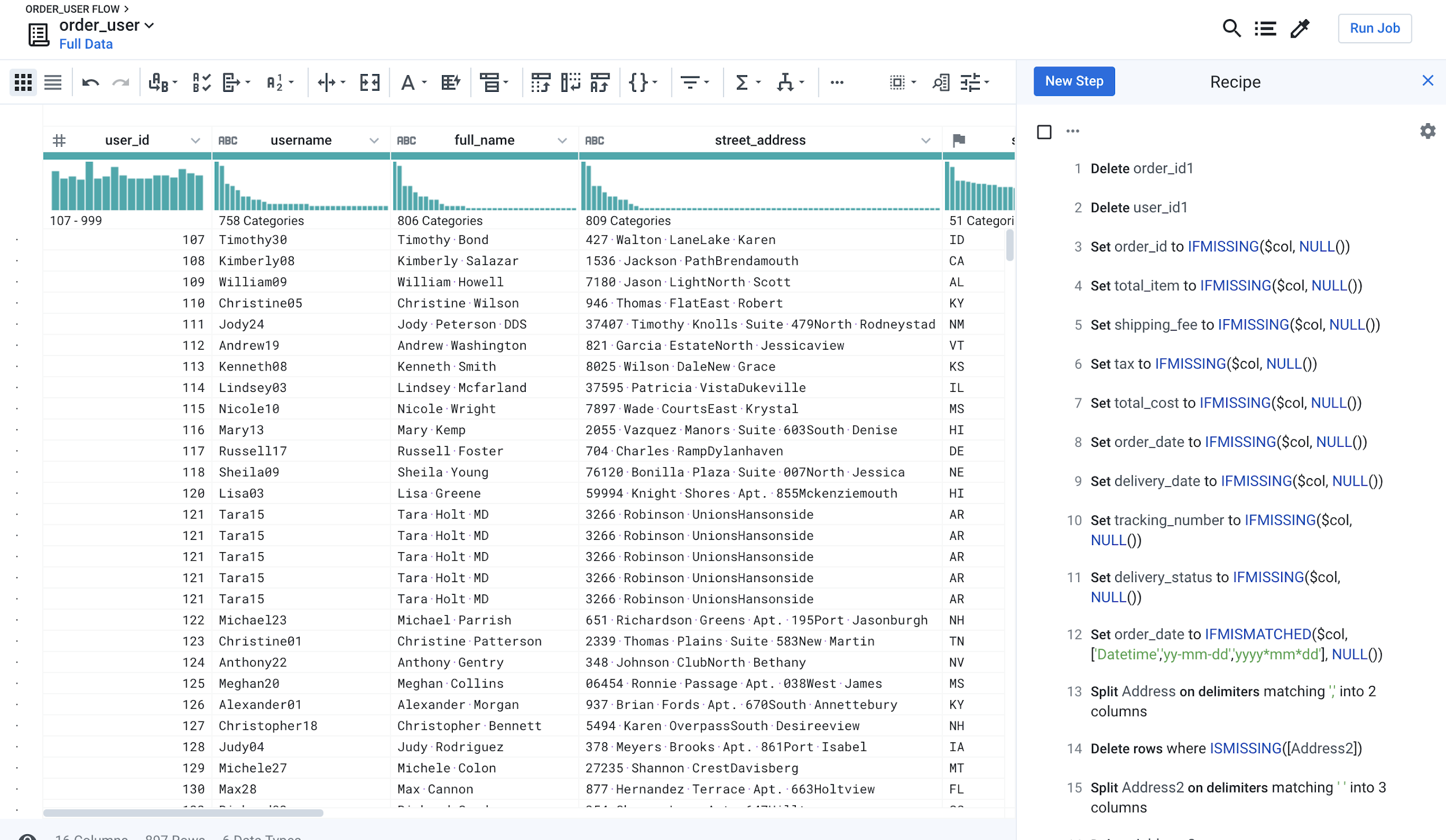
It is an Intelligent Platform that Interoperates with Your Data Investments. It sits between the data storage and processing environments and the visualization, statistical or machine learning tools used downstream. Trifacta is a tool in the **Big Data Tools** category of a tech stack.

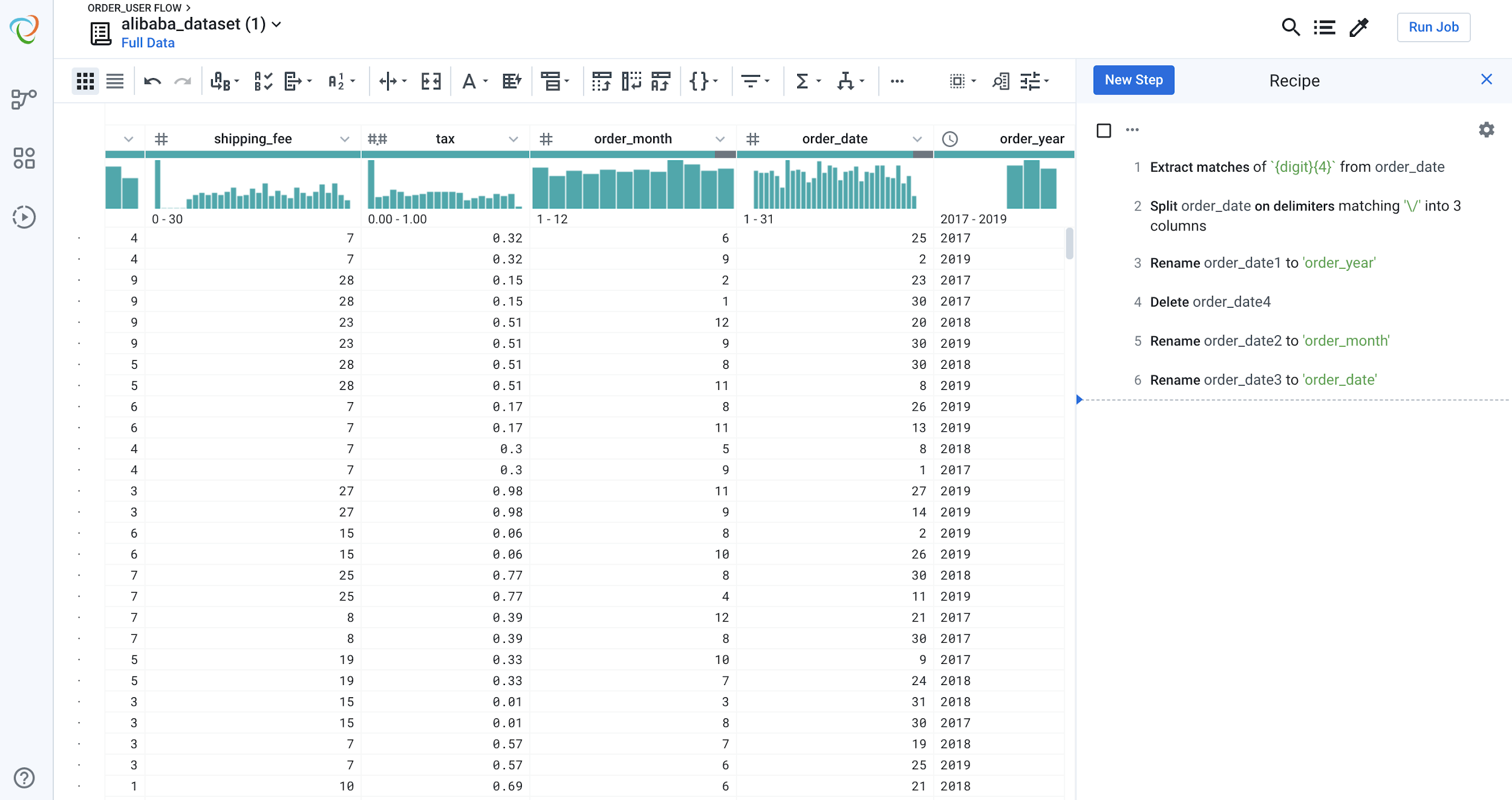
## How we used Trifacta for our datasets?

**We used Trifacta to split columns , clean the data , fill in N/A values, remove mismatched values**

## What are the pros of Trifacta?

* Interactive Exploration
* Automated visual representations of data based upon its content in the most compelling visual profile
* Predictive Transformation
* Intelligent Execution
* Collaborative Data Governance





# Salesforce Einstein Analytics

Einstein Analytics is an app used to visualize the activity occurring in your Salesforce environment. Whether you use Salesforce for Sales, Marketing, or Service, this visibility tool offers insights into the data (like contacts, campaigns, or accounts) your users add to the CRM every day

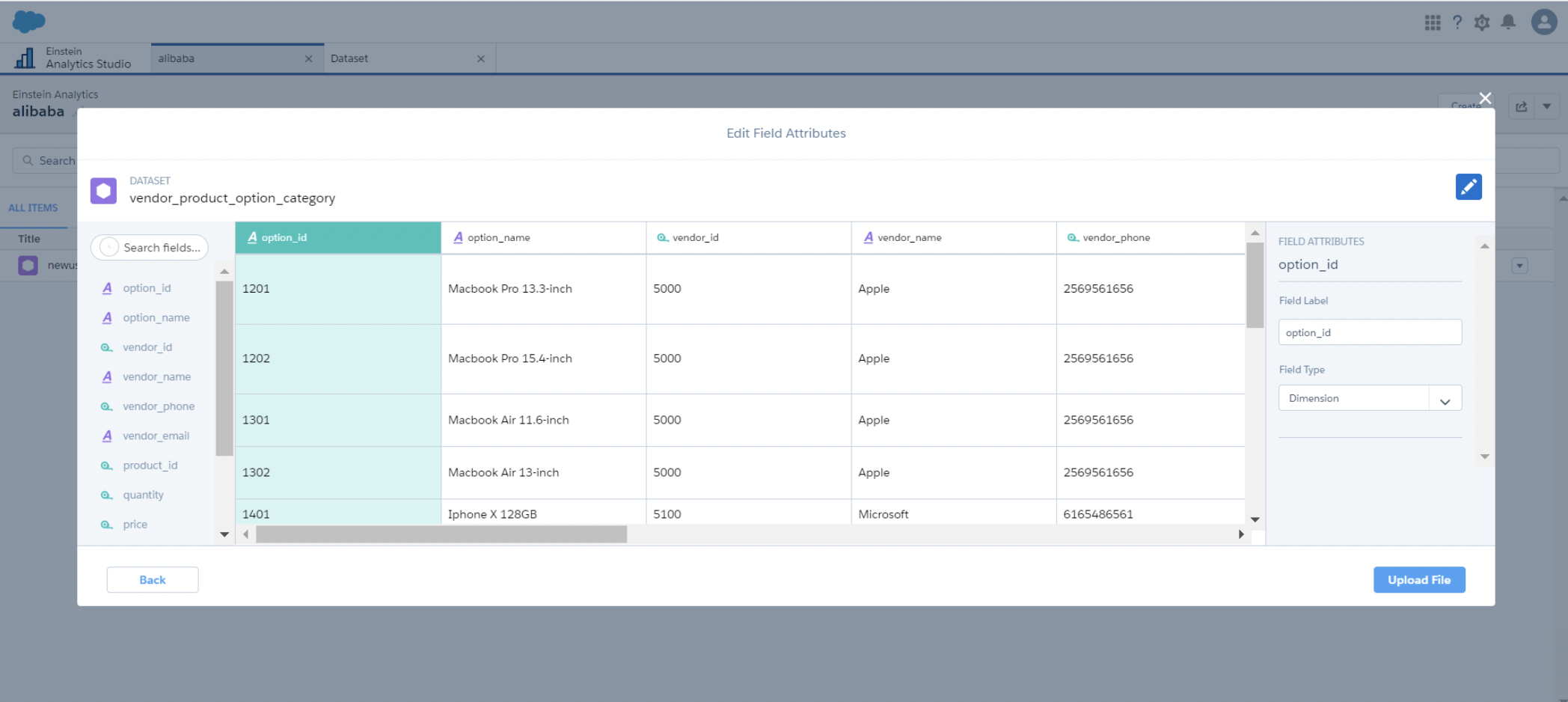
## How we used Einstein Analytics with our datasets?

**We built a dataflow joining our two main tables Users and Vendors using Einstein Analytics. We used it to build our dashboards to gain insights into our dataset**

## What are the pros of Einstein Analytics?

* Easy visualization tool for creating dashboards
* Intelligent prediction builder
* Smart action recommendations
* Easy access to data science features without the actual need of coding

## Which columns are dimensions, which columns are measures?



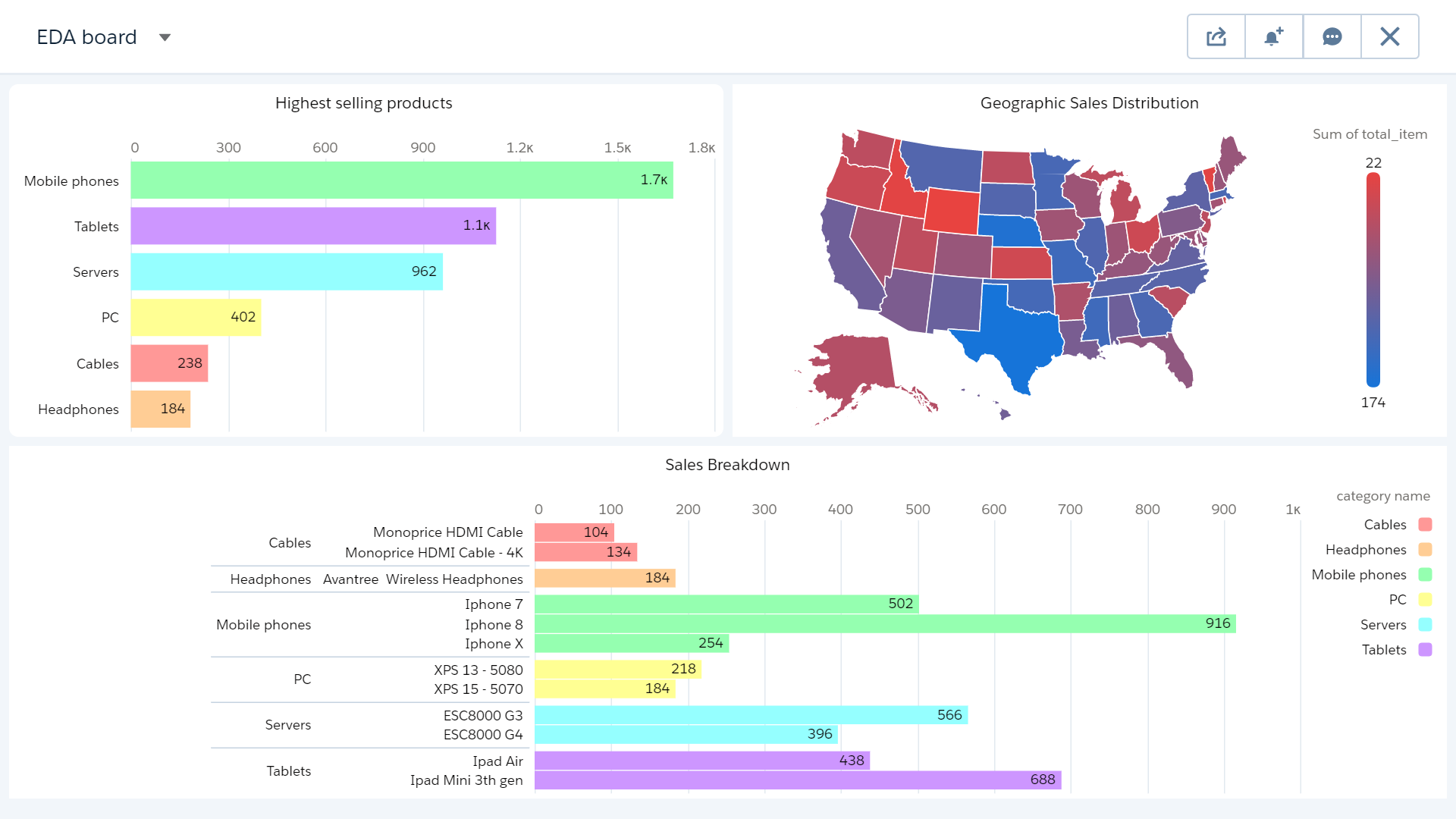
## Data flow created to join the ‘User’ & ‘Vendor’ into one table



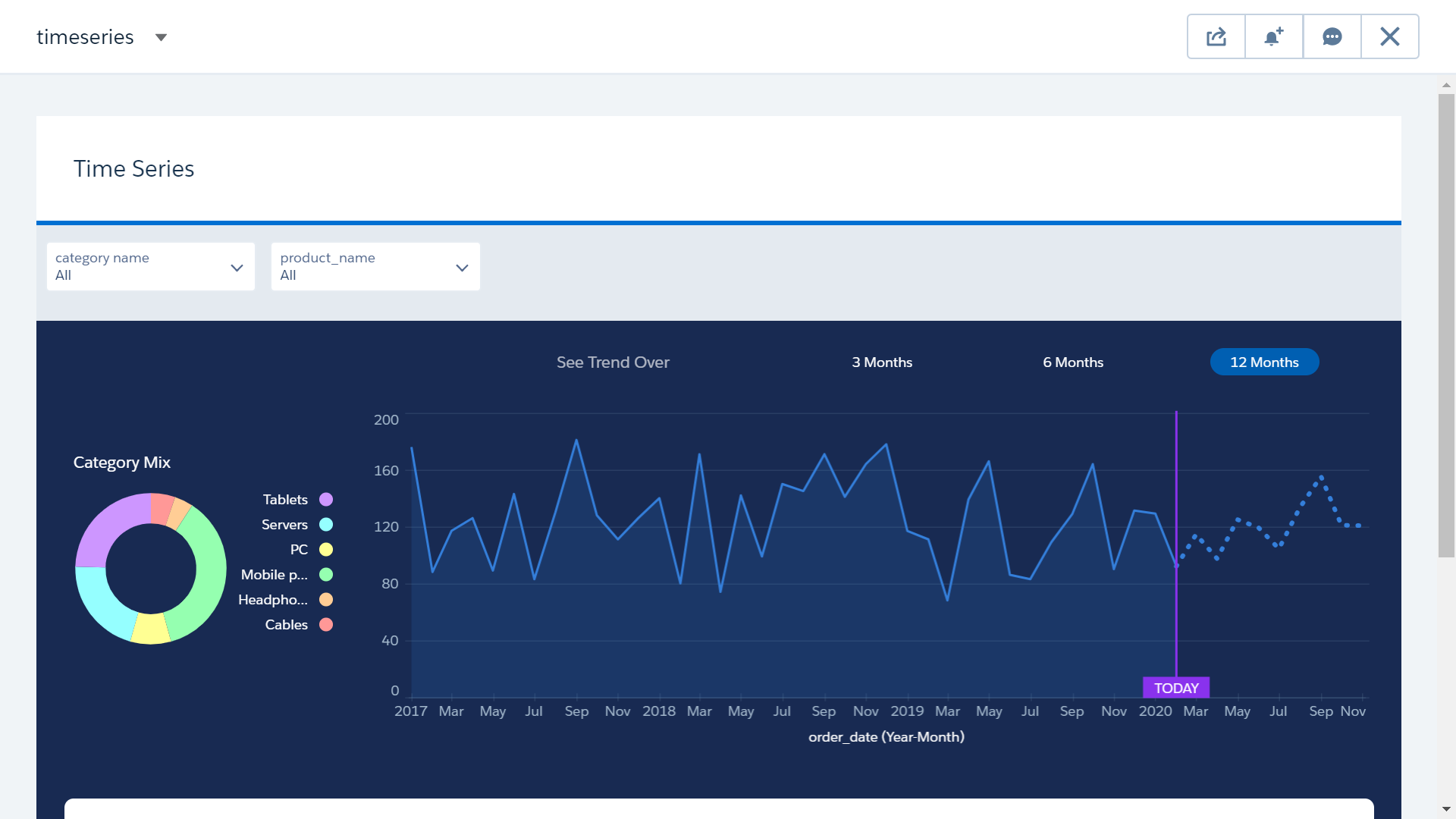
## Who would use this dashboard?

**Sales Insights**

Sales Insights arms your sales team with the data needed to make informed communication with a potential customer and generate revenue faster. The sales & marketing team can use this dashboards as a proof-of-concept to use the right marketing strategies to increase the revenue of their company.



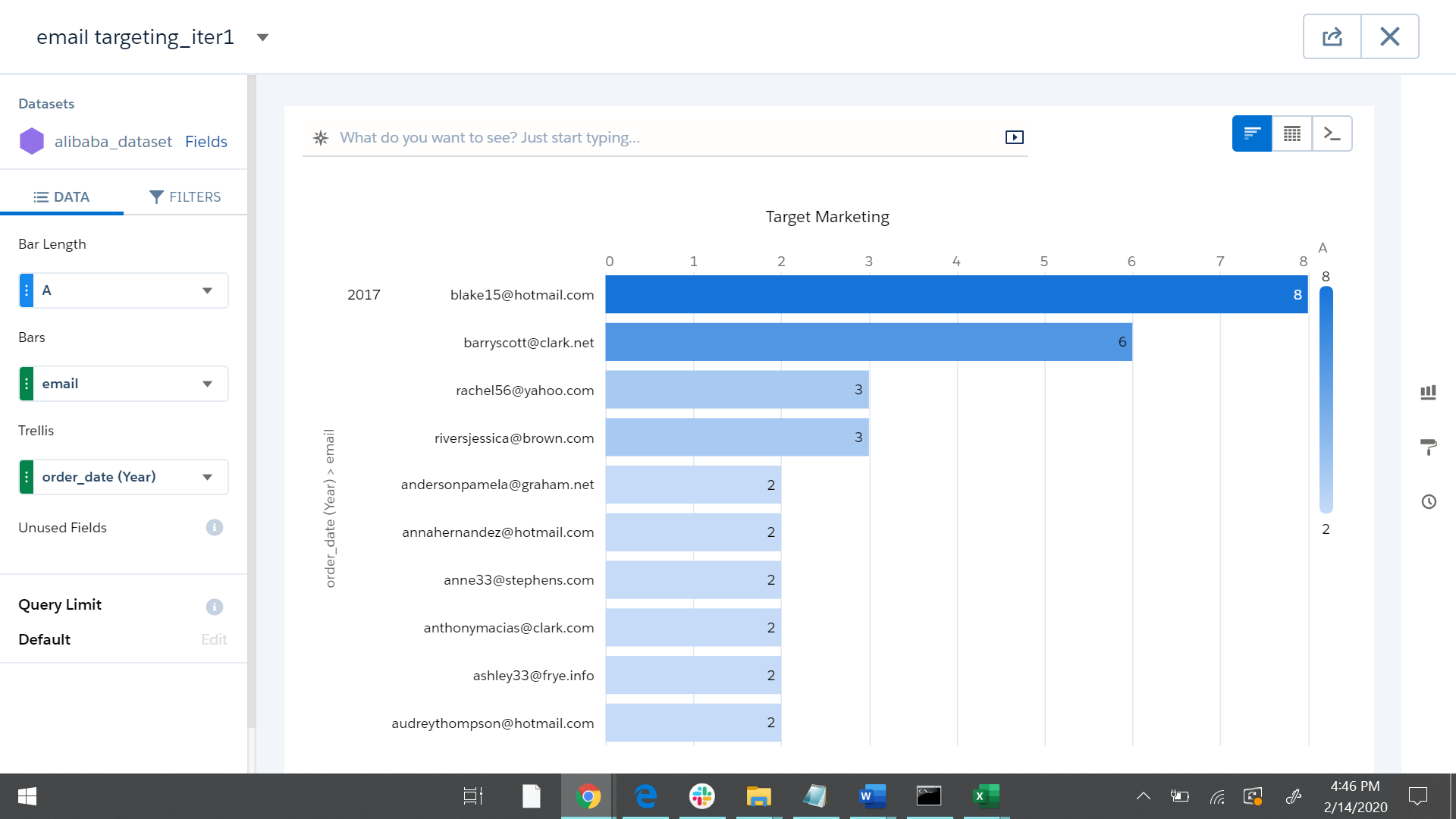
The below dashboard is part of Sales Insights which gives us a Time series forecasting for the upcoming year



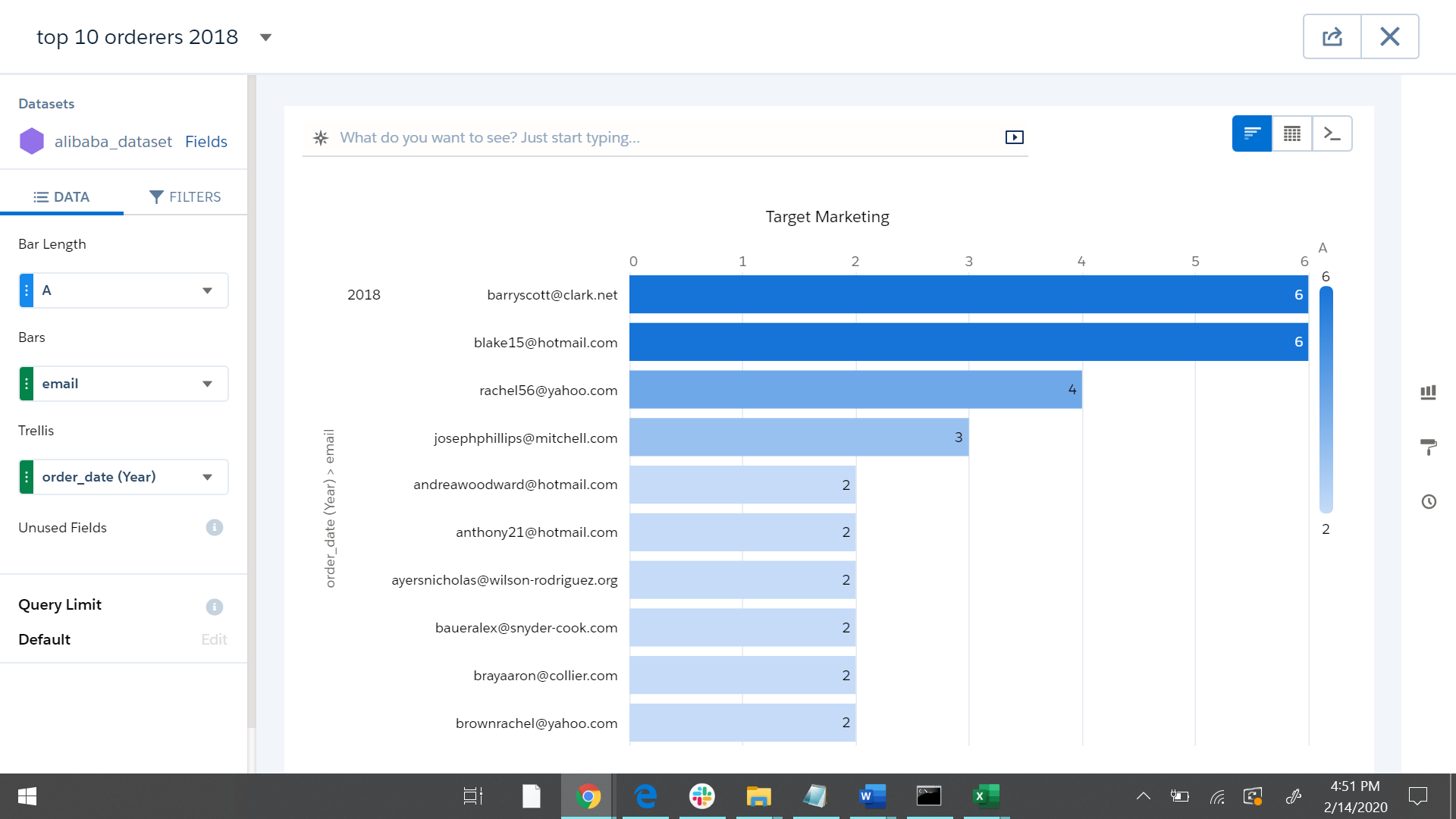
**Email Marketing**

Marketa can send targeted marketing and email campaigns to further engage their prospects called Smart Campaigns. Tracking data and metrics allow you to see which components are the most successful and which lead to higher conversion rates to provide **Promotions , Deals and Discounts** to their customers.

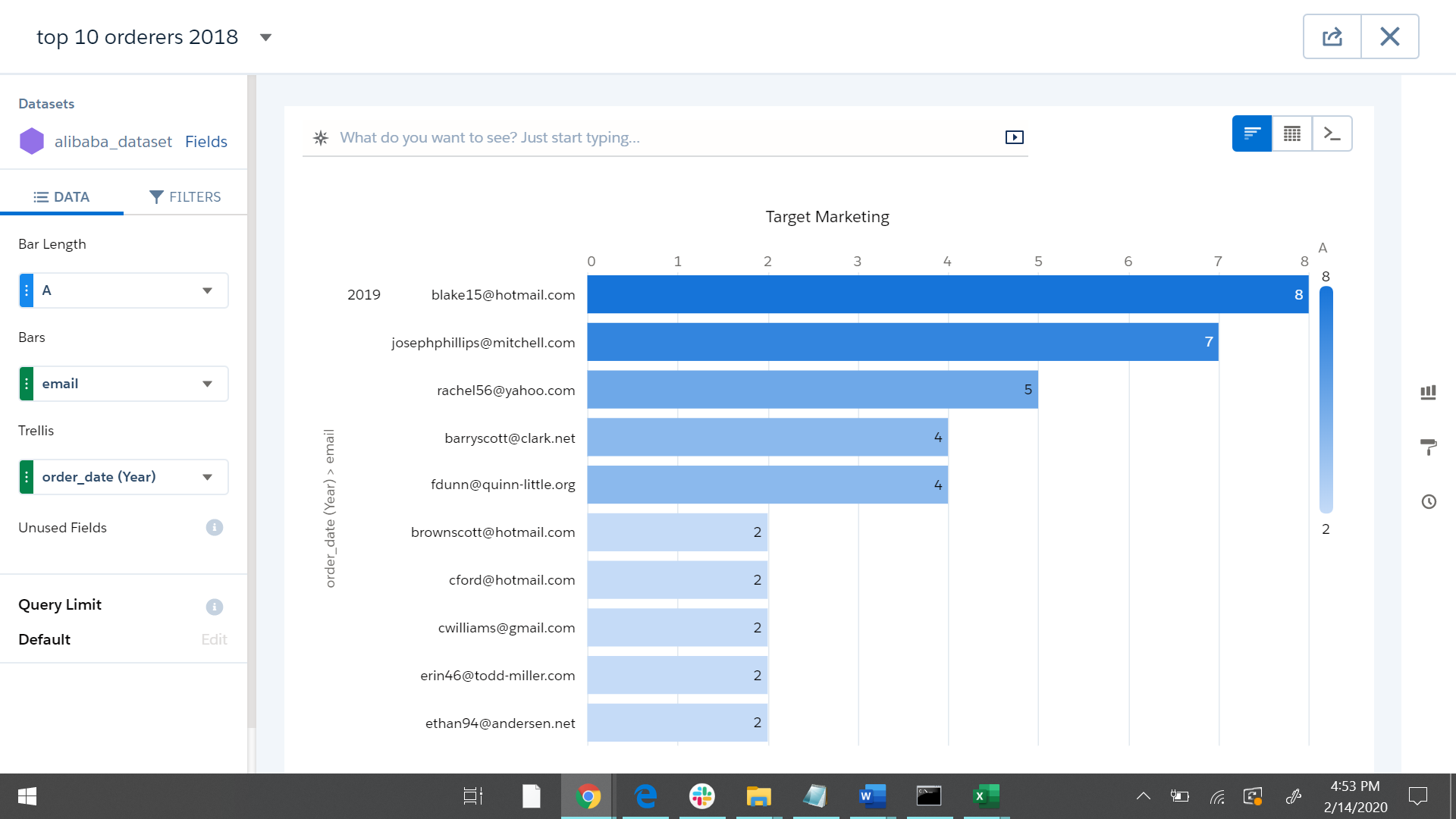
**Top 10 active customers for 2017**



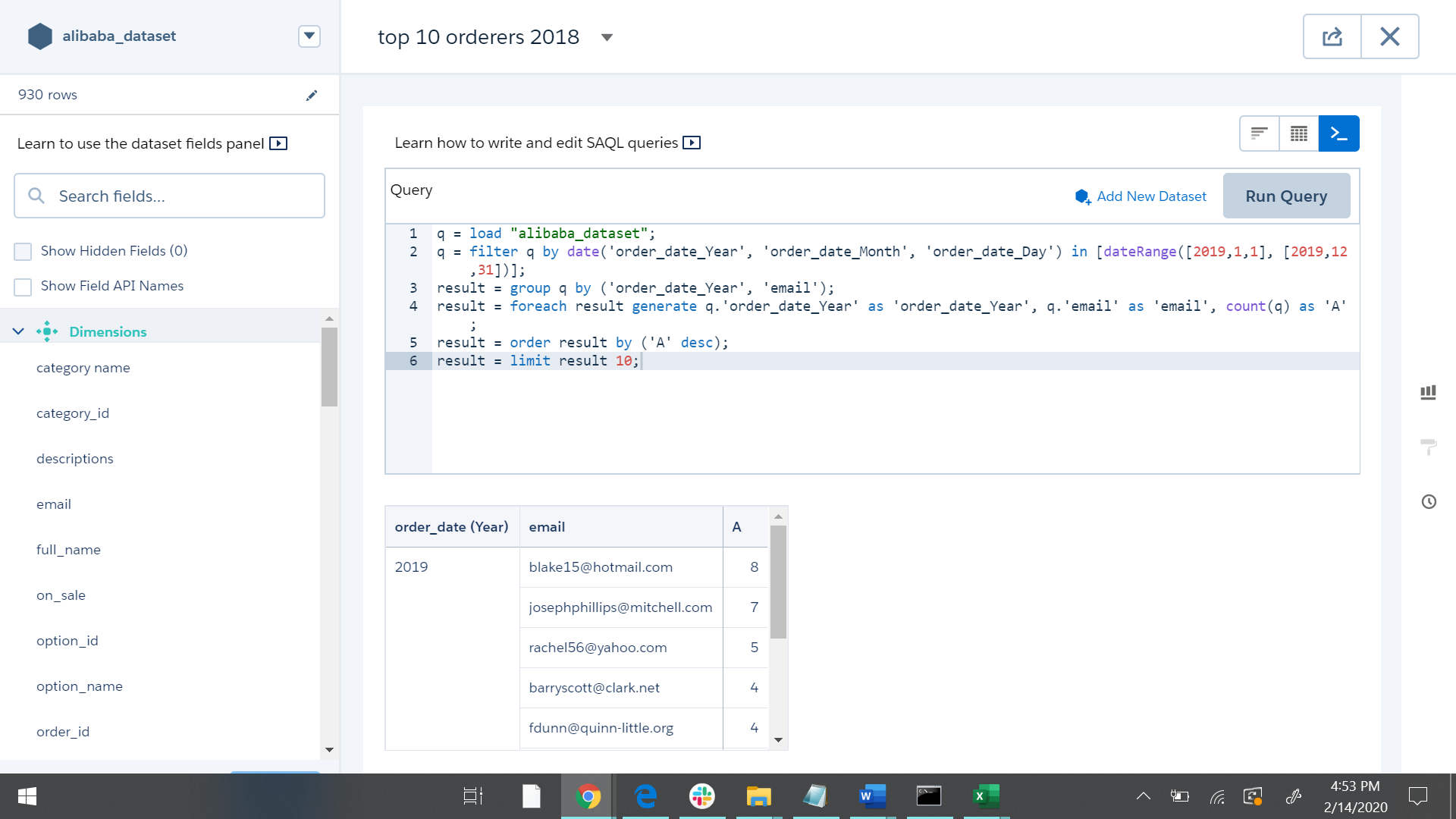
**Top 10 active customers for 2018**



**Top 10 active customers for 2019**

****

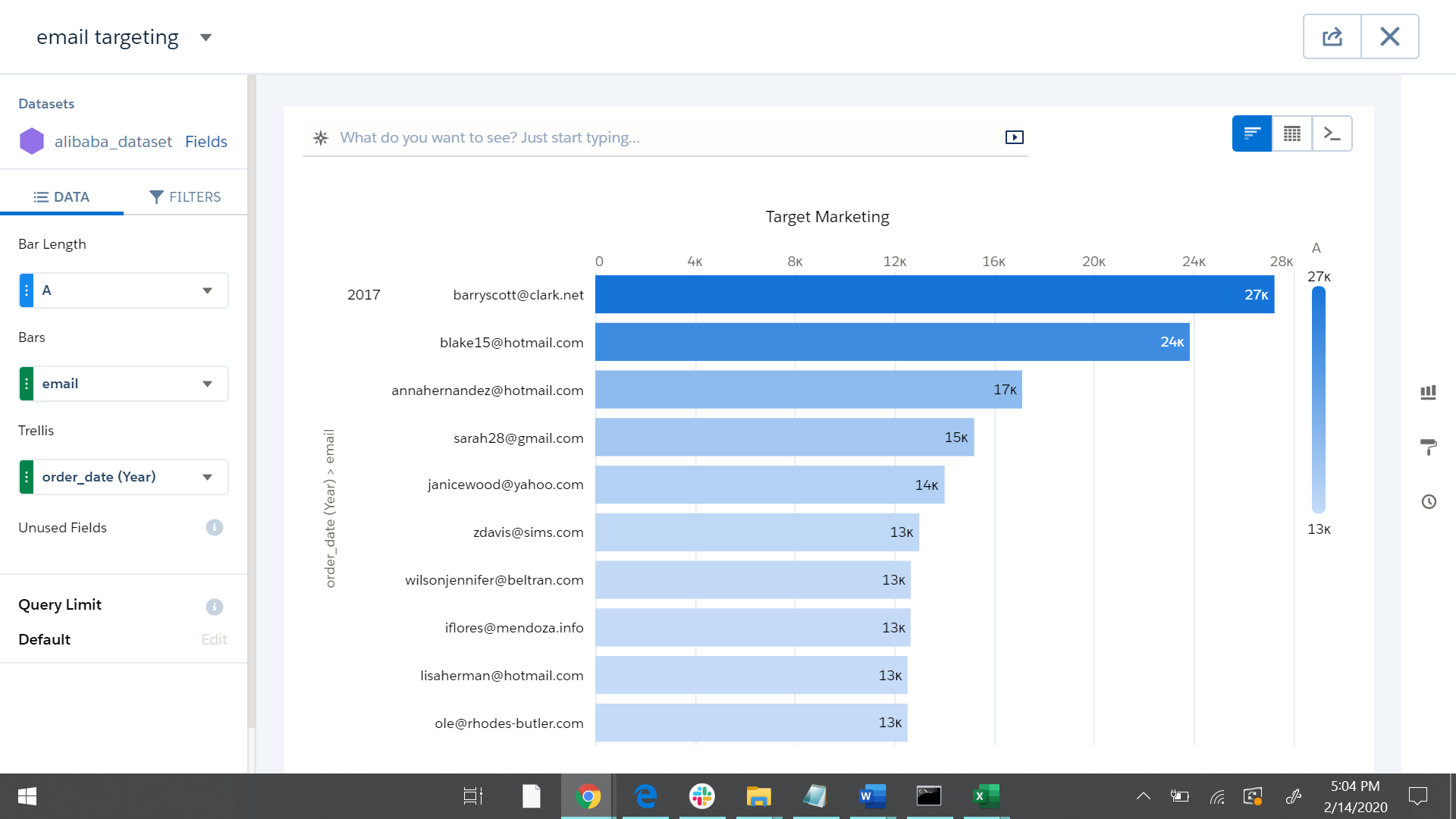
**SAQL query for retrieving the Top 10 buyers**

****

**Recommendations and Promotions for the Top 10 customers**

Marketa can offer customized recommendations and promotions for these customers which would help customer retention

**Top 10 Big Spenders for 2017**

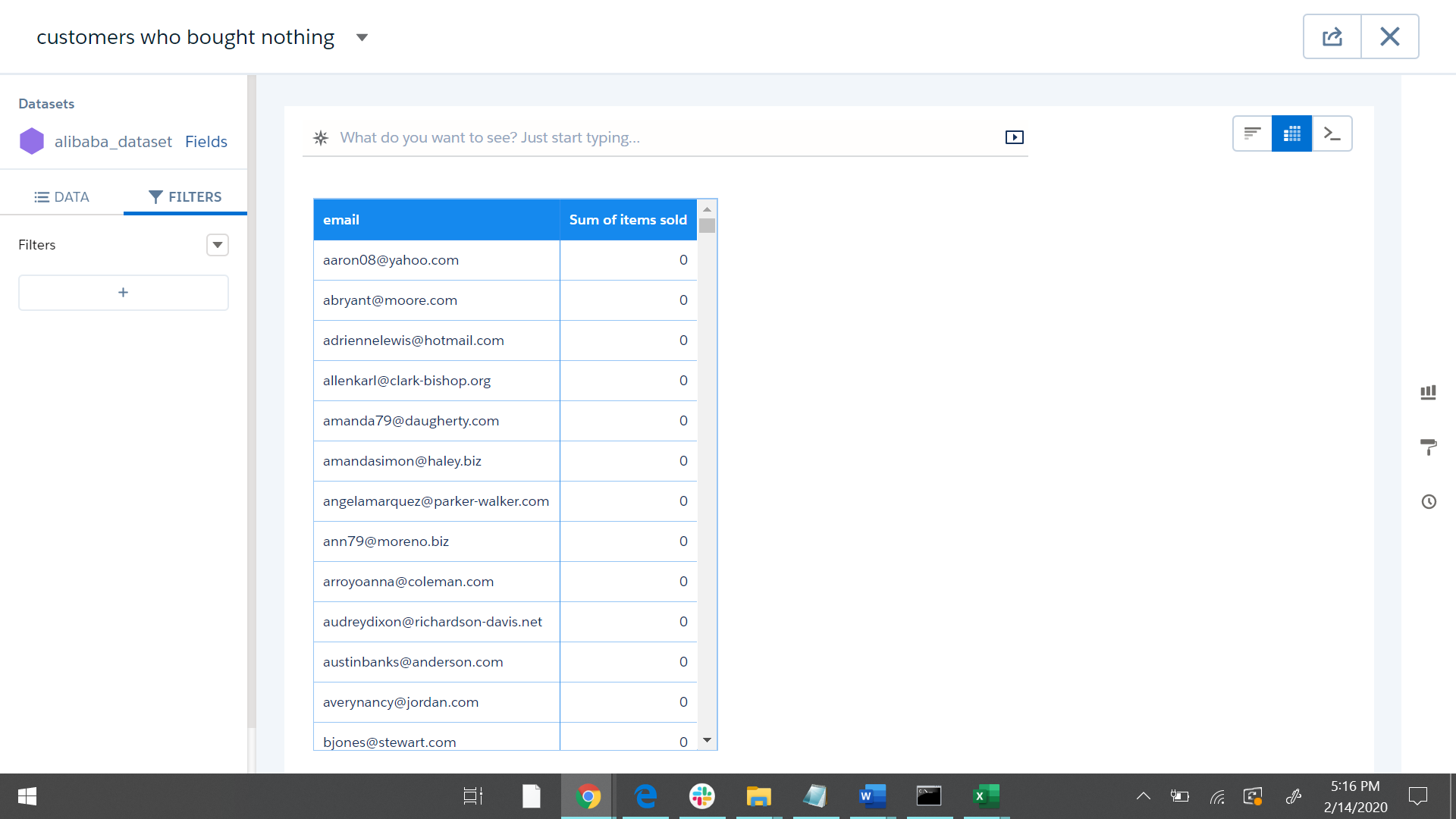
****

**Recommendation for Top 10 Spenders**

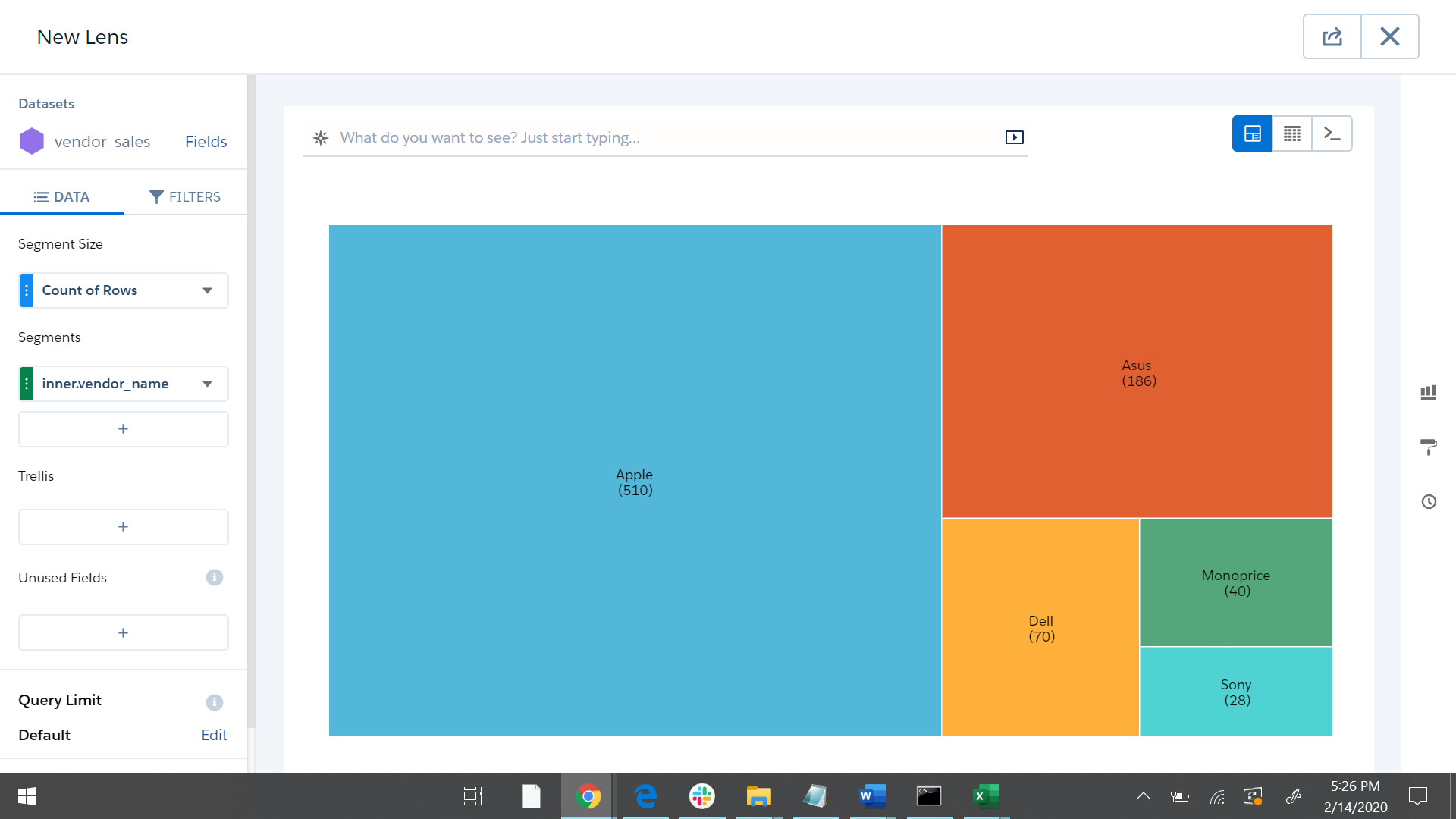
Marketa can offer deals to the customers who are spending more on their website

**No Buyers**

Marketa can target these customers as they have not bought products from 3 years

**Top selling vendors**

Marketa can focus on the top selling vendors so that they can invest more on these vendors



# Future Scope

* Marketa can hire a third-party to track the customers transactions through their credit card details
* Marketa can analyze cart abandoned rate as a metric to target their lost customers
* Marketa can track user sessions to reduce the customer’s bounce rate

|  |
| --- |
| Reference links :  <https://www.trifacta.com/>  <https://github.com/tshinohara/xsv>  <https://pandas.pydata.org/>  <https://www.geeksforgeeks.org/python-faker-library/>  THE END |