Lab 15. Preparing Data for Analysis with Tableau Prep

This lab will cover the following recipes:

- Installing Tableau Prep
- Building the first flow with Tableau Prep
- Working with big data

Introduction

The main role of any business is bringing value to its customers, employees, and stakeholders. Every group needs to own a unique value, for example, stakeholders are interested in revenue growth and meeting KPIs. On the other hand, customers are expecting high-quality services and products.

In order to drive value, businesses need managers and employees to make the right decisions. They require clean and accurate data that enables sound decision-making and allows them to generate value for everyone.

Modern businesses generate tremendous volumes of data. Some data is available in data warehouses, some in [**online Transaction processing**] ([**OLTP**]) system, and some in third-party marketing applications. Tableau is a powerful tool that allows us to explore and visualize all this data, but sometimes it isn't enough. In some cases, data should be clean and transformed before analysis and usually it requires the help of data engineers or [**extract,transform,**] [**load**] ([**ETL**]) developers. As a result, this is a kind of bottleneck for the organization and slows down decision-making processes as well as value-generation.

With Tableau 2018.1, a new desktop tool - Tableau Prep- was announced. This is a self-service data tool that will help clean, transform, and reshape data for better analysis. It empowers business users with the ability to see and feel data and visually manipulate it in order to shape it into the correct form.

Installing Tableau Prep

In the modern world, analysts and business users don't want to be block with IT any more. In other words, in traditional organizations all data work is done by IT stuff and business users have to wait, while IT solve their ticket. They demand data; they want to get all the data and analyze it in order to create vital insights that will help them to survive in this highly-competitive world. Around the globe, many people have adopted Tableau and they use for their day-to-day tasks.

Business users are becoming more and more proficient with technology and data. They are ready to learn new skills that help them get faster insights. As a result, Tableau released a new tool: Tableau Prep. It's an efficient and powerful desktop tool that's available for Windows and macOS and has a rich functionality for shaping data.

You might get questions about the use cases for Tableau Prep. There are lot of use cases for this tool. The main benefit of Tableau Desktop is that it empowers end users by giving them a powerful data-exploration tool. In most cases, business users with Tableau don't depend on IT any more. But they still need IT guidance when they want to prepare their data for analysis. By releasing Tableau Prep, Tableau is trying to solve one more challenge for business users, and offers them rich capabilities for local data transformation and preparation without IT involvement. For example, in a marketing team, you might have lots of data sources, and every month you need to bring in new data sources, so you have to move fast. Using Tableau Prep, you can set your own flow and bring all the data together. You can join, transform, reshape and clean your data, and generate a Tableau data source that will be ready for data analysis and exploration.

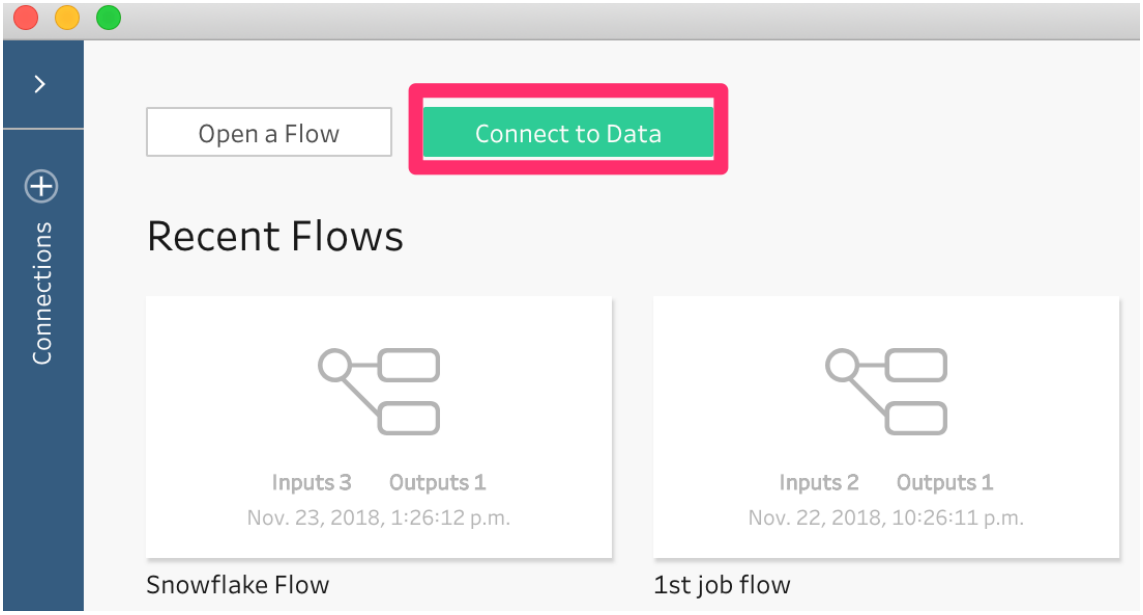
Getting ready

First of all, we should download Tableau Prep from the Tableau website and install it on our machine. All Tableau software releases can be downloaded from https://www.tableau.com/support/releases/.

How to do it...

Let's find Tableau Prep, download it, and install:

1. Go to https://www.tableau.com/support/releases/prep.
2. Find the most recent release of Tableau Prep and download it for your OS.
3. Install Tableau Prep and launch it:



How it works...

Tableau Prep has a basic interface. You can discover `Sample Flows` or open your `Recent Flows` section. Flow means a sequence of steps (data transformations). Basically, we can think about Tableau Prep as a desktop ETL tool that allows us to connect and extract data, transform it, and publish it into a Tableau data source or write it to a file.

Currently, Tableau Prep supports fewer data sources than Tableau Desktop, but it offers over 40 different data sources, including Snowflake, an innovative analytical data warehouse.

There's more...

When we download Tableau Prep for the first time, we can start a free trial and we can use Tableau Prep for 14 days. Then, if we want to continue to use it, we need to buy a license. Tableau Prep isn't an individual product, it comes with Tableau Creator License and includes Tableau Desktop and one Tableau Server or Tableau Online. You can read more about licensing costs here: https://www.tableau.com/pricing/individual.

Building the first flow with Tableau Prep

After successfully installing and launching Tableau Prep, we can start to build our first data flow using the sample dataset. (Just an example, in this recipe, we'll cover how to connect the data, transfer it, and then publish the result.)

Getting ready

To proceed with this section, download the Microsoft Excel document, `installs.xlsx`, that's available for this lab.

This dataset has data about a number of app installs for iOS and Android by date.

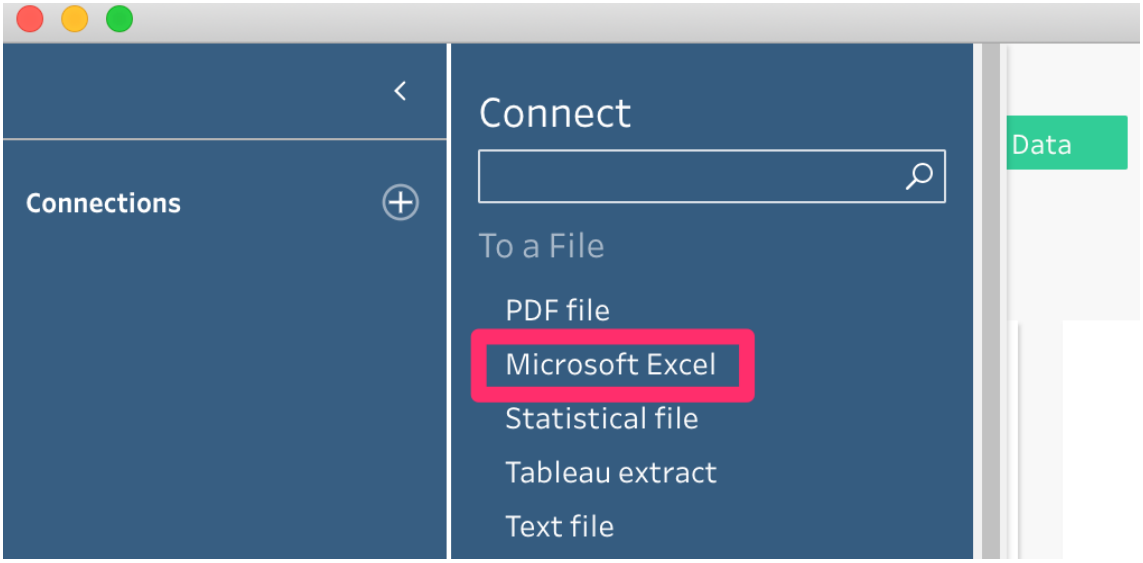
How to do it..

In the following section, we'll look at how to connect the data, transform it, and then publish the result.

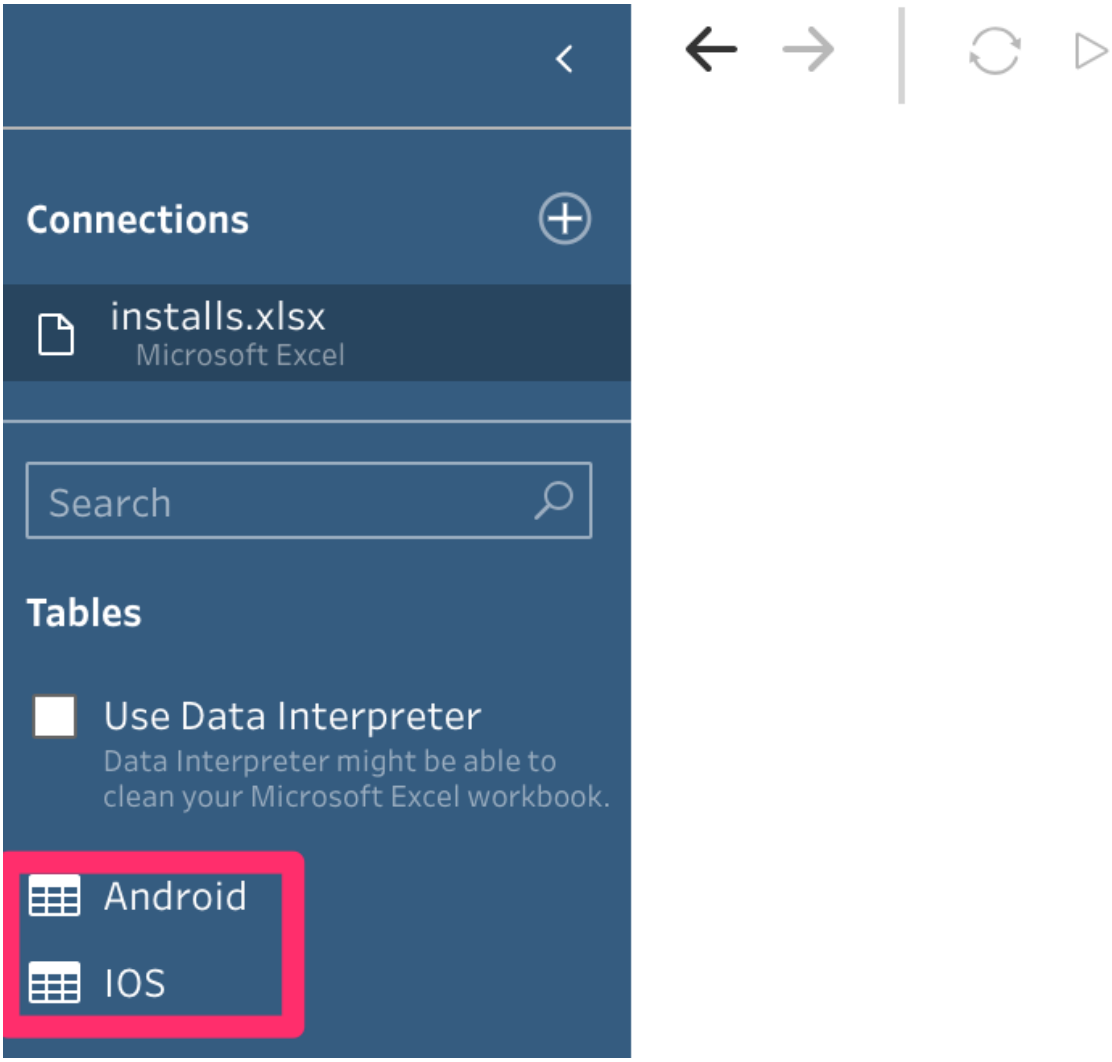
Connecting the data

Tableau Prep supports lots of data sources and we can easily blend together multiple data sources, such as files and databases. Let's get started:

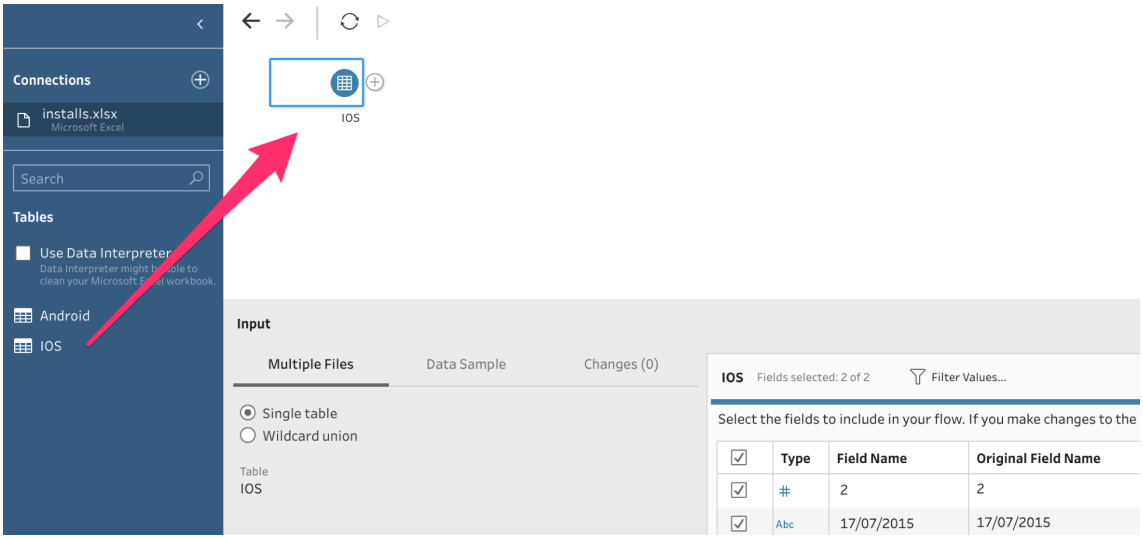
1. Navigate to **`File`** | **`New`** and create a new flow.
2. Click on the plus + sign near **`Connections`** and choose **`Microsoft Excel`** :``



3. Choose the `installs.xlsx` file and click on **`Open`** . Connect the file and it will show us two available tables, **`Android`** and **`IOS`** , as shown in the following screenshot:``



4. Using the drag-and-drop method, drag `IOS` to the canvas. Tableau Prep will create the `Input` step.



During the Input step, Tableau will read the file and learn about the structure of the file. For example, you can use the

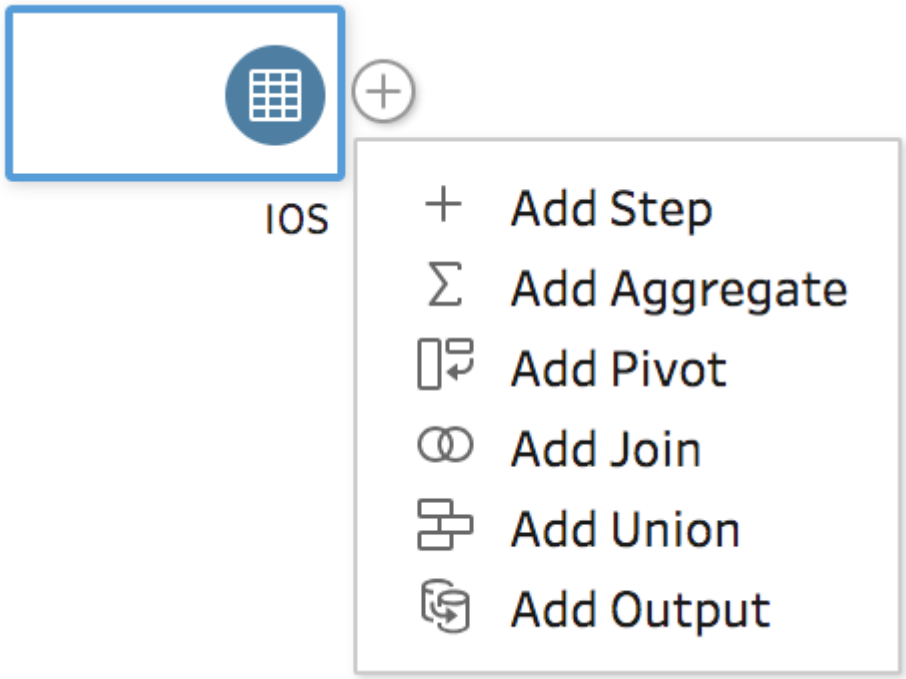
CSV or TSV file, and Tableau will automatically recognize the pattern and try to split into columns. We can also use the `UNION` operation by using the Wildcard union in order to read all the files and merge them into one. In addition, we can sample data in the case of a big dataset. Finally, we can apply filters.

5. Let's add one more dataset file with the Android `installs`. We have two ways of doing this: we can either drag and drop the **`Android`** sheet to the canvas or adjust the settings of **`IOS`** by enabling the Wildcard union.
6. We'll use the first method. Drag and drop **`Android`** to the canvas.

Transforming the data

Once we've successfully connected the data and added data sources to the canvas, we can start to build our flow:

1. Click on the + sign near the data source to choose the next step:

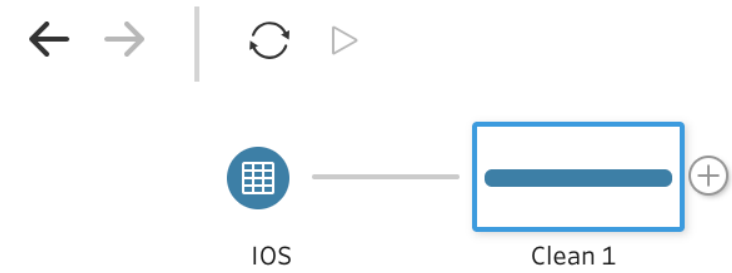


We have the following options:

Steps	Description
`Add Step`	This step will allow us to look at the data and modify it.
`Add Aggregate`	We can calculate a new measure, using functions such as `SUM` , `AVG` , and `COUNT` .
`Add Pivot`	We can transpose columns into rows. In other words, we can convert cross-table into a normal table.
`Add Join`	We can join data streams using the `INNER` , `LEFT` , `RIGHT` , and `OUTER` joins. Moreover, Tableau will visualize and color results on the fly.
`Add Union`	We can merge multiple streams into one.
`Add Output`	This is the final step, where we'll generate the result set. We can write into the CSV file or the Tableau data source. In addition, we can publish directly to the Tableau Server.

In our case, we will add the **`Clean`** step and learn about our dataset.

2. Click on the new step, **`Clean 1`** , and explore the **`Profile`** pane, as shown in the following screenshot:



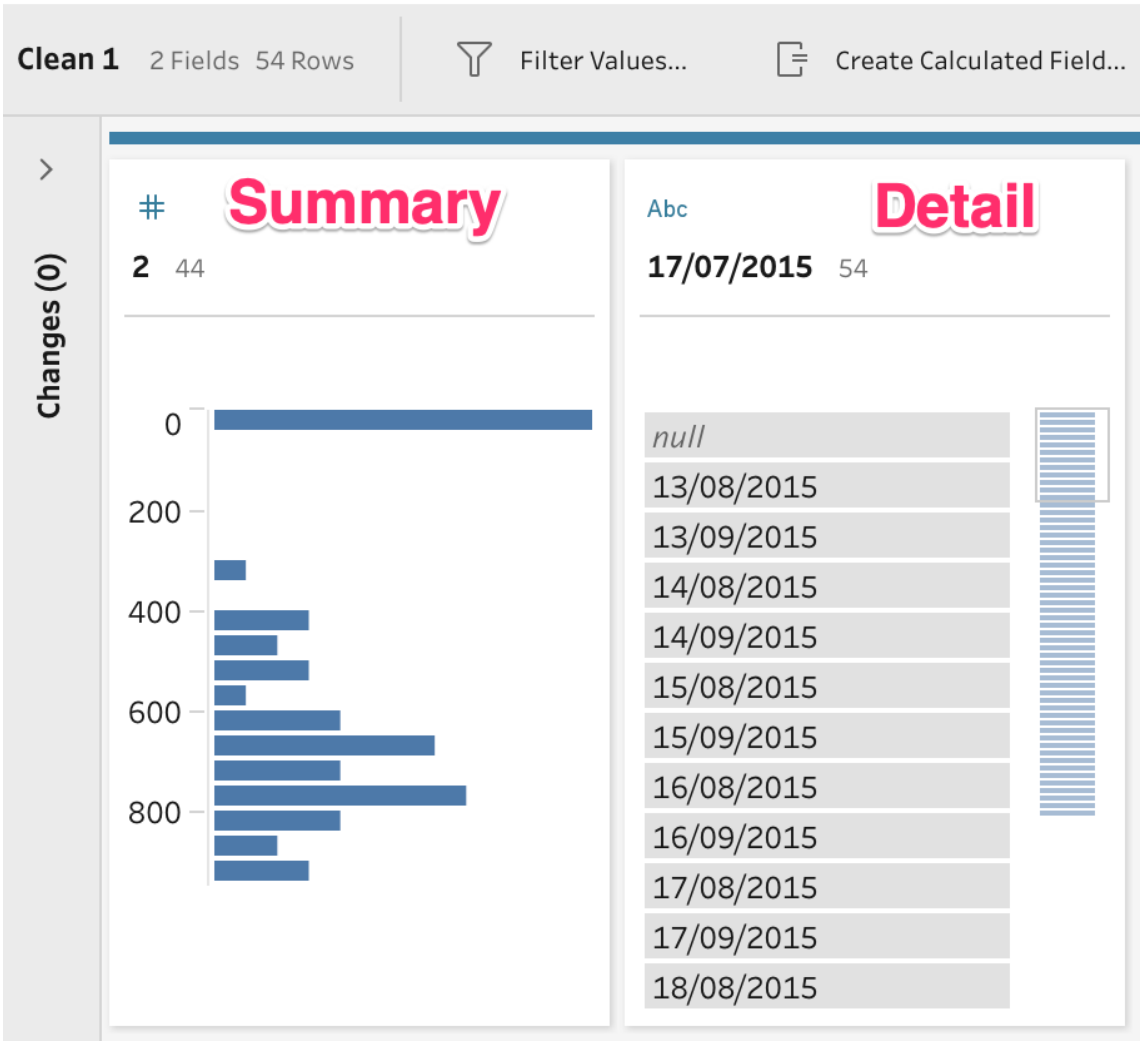
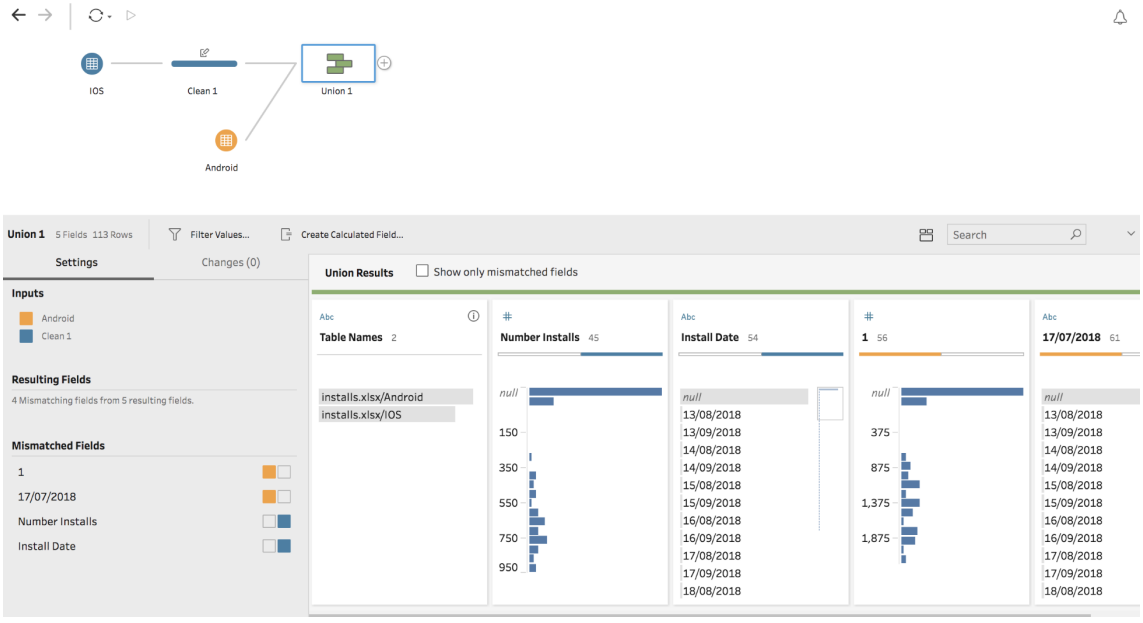


Tableau Prep creates cards for each column. At this step, we can modify our data by changing the data type, creating a new column, and filtering values. In addition, we can see the connection between values. For example, if we click to any value in the data card, it'll show us dependent values. Moreover, it'll show us a histogram of value distributions. Data card can have two [**View States**]--- `Detail` (column #2) and `Summary` (column #1), that, is distinct values or grouped-by values.

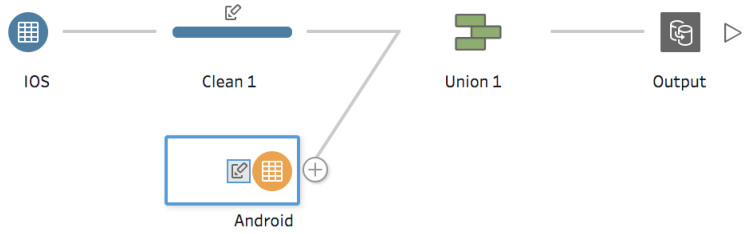
3. Rename the columns. The first column should be `Number Installs` and the second should be `Install Date`.

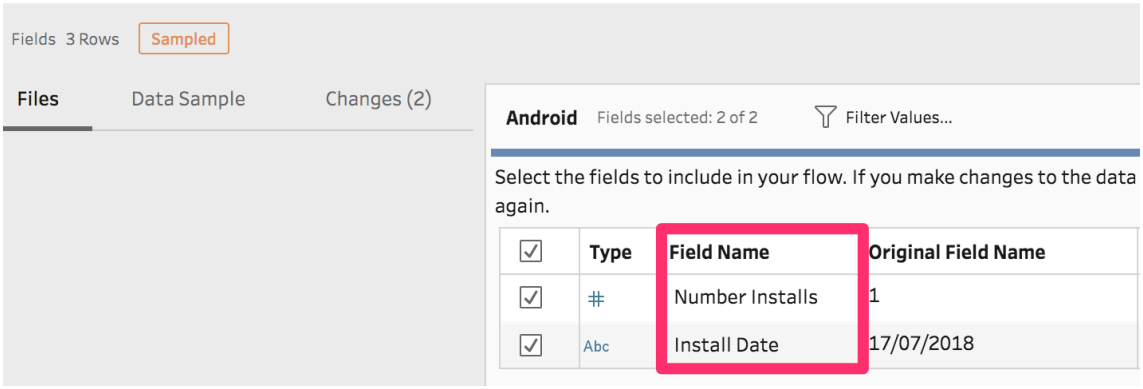
4. Union the datasets. Click `+` after the `Clean 1` step and choose `Union`. Drag the `Android` data source and drop it on top of the `Union` step. At the same time, we can add [**Union 1 step**], when we drag and drop the `Android` dataset on top of the `Clean` step, it will ask us to choose the appropriate step, such as `Join` or `Union`:



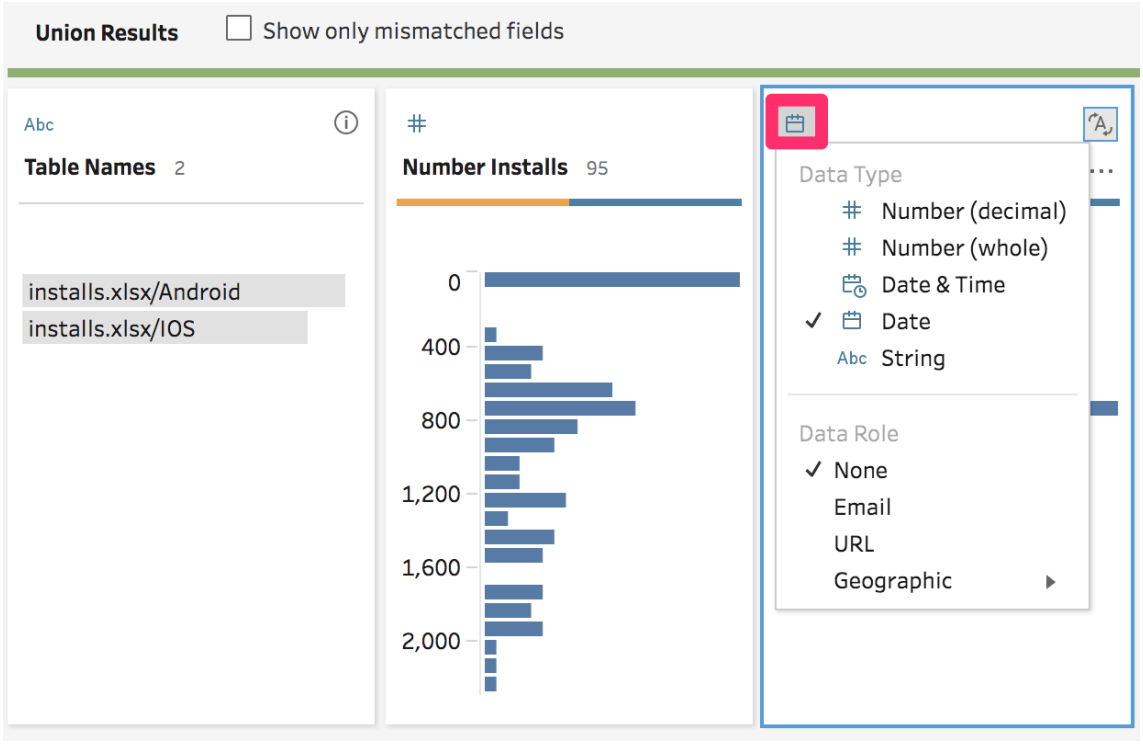
You will see that the result set doesn't look good in the `Profile` pane because we didn't rename the `Android` dataset column names and you can see the orange (Android) data still in different columns. Moreover, our date is in string format.

5. Click on the `Android` dataset and rename the column names the same way we did for `IOS`. Click on the `Android` Data Source Input and rename both `Field Name` to the `Install Date` and `Number Installs accordingly`:

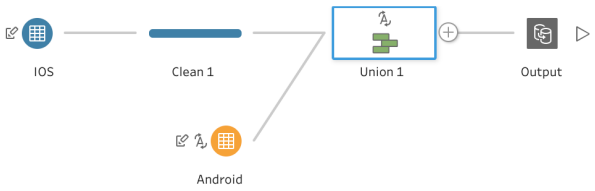


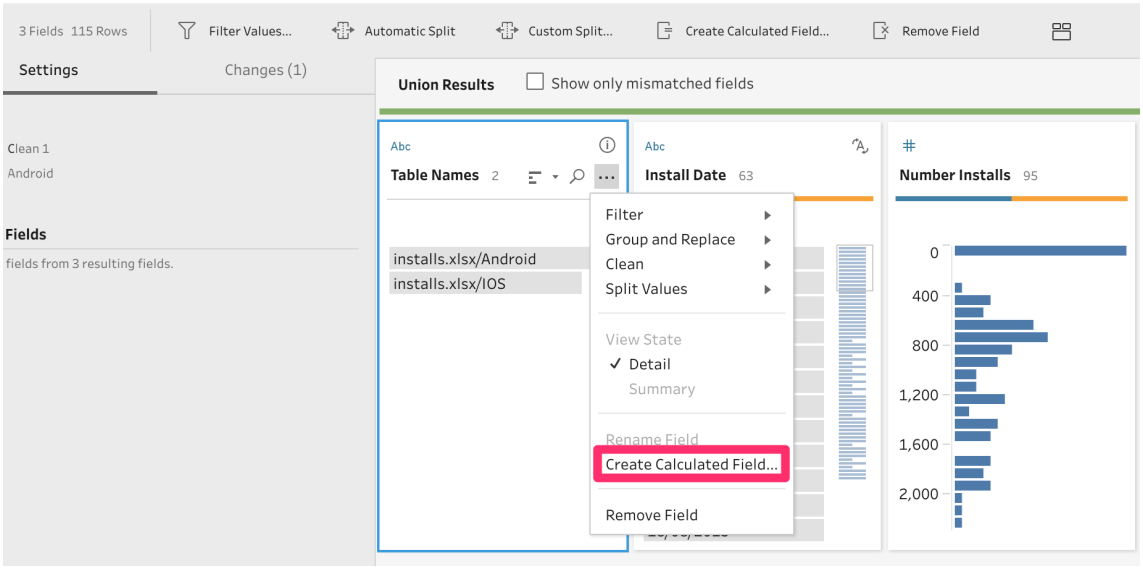


6. Click on the `Union` step and change the date format:

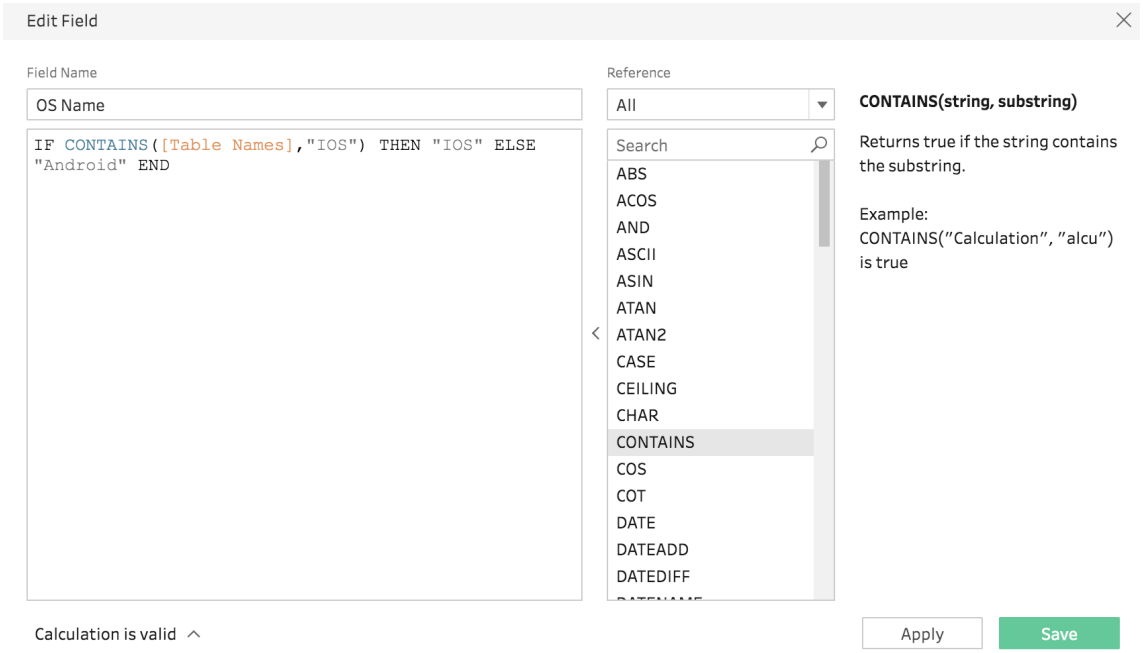


7. Let's also create new calculated fields. We don't have an OS name. We can use the [**Table Names**] system field in order to extract the phone's OS name. In order to create a new calculated field. In order to do it, click on the [**Union 1**] step and click on the [**Table Names**] pane and choose [**Create Calculated Field...**]:





Then you can write this statement. Tableau Prep has the same syntax as Tableau Desktop.



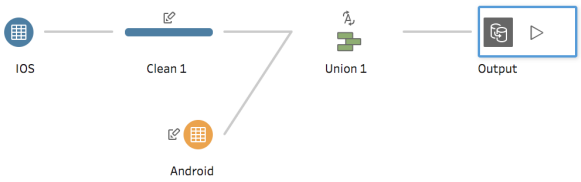
This will add one more field to our dataset. Now, we need to drop the original **`Table Name`** field. The calculated field won't disappear. In order to drop field, you should click on **`Remove Field`** at Data Pane of this field.

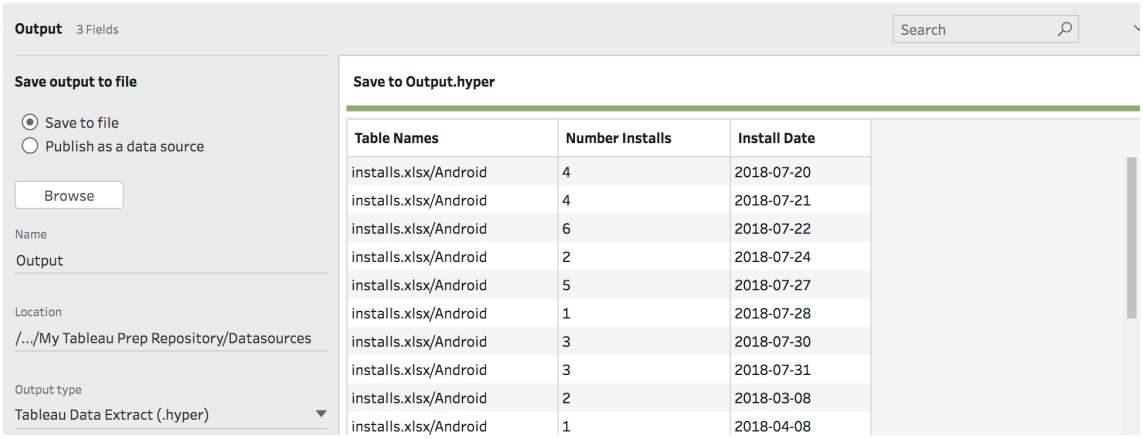
As a result, we've got the correct dataset for further analysis in Tableau.

Publishing the result

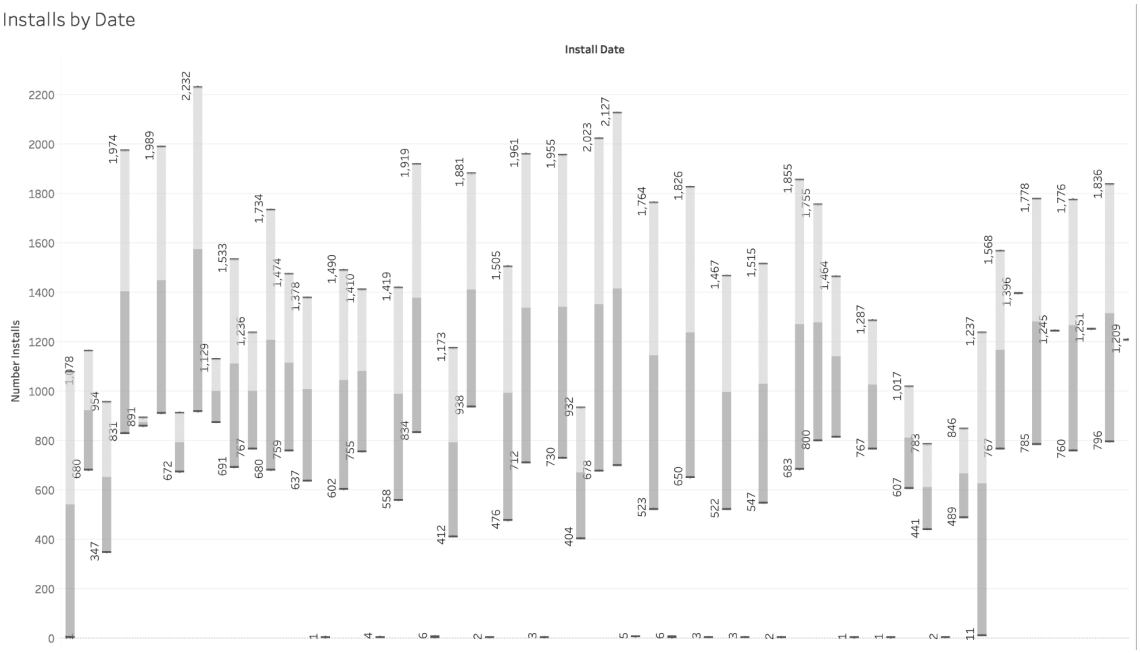
When we finish the transformation, we should publish our result using the `Add Output` step. Let's create the Tableau data source:

1. Click the `+` sign and choose `Add Output`. We have the option to save to a CSV file or to create Tableau Data Extract. Moreover, we can publish our data source right to the Tableau Server.
2. Create a Hyper data extract:





3. Click on the `Run Flow` button, this will find a new Tableau data source.
4. Open the new data source with Tableau Desktop:



How it works...

Despite the fact that we performed a very simple task, we covered 80% of Tableau Prep's functionality. The main goal was to understand how the product works. Basically, Tableau Prep can connect to any data source and gives us the ability to fully control the data flow by transforming and merging the data. In addition, we can create calculated fields using the Tableau syntax. Finally, we store our results in CSV or Tableau Data Source.

There's more...

In this recipe, we got introduced Tableau Prep and learned the tool's main functionality. Here's a good resource for doing more complex work with Tableau Prep: https://onlinehelp.tableau.com/current/prep/en-us/prep_dayinlife.htm.

It has the following two use cases with a detailed step-by-step guide:

- [*Hospital Bed Use with Tableau Prep*]
- [*Finding the Second Date with Tableau Prep*]

Working with big data

Tableau Prep works with big data volumes and big data tools, such as Snowflake, Redshift, and Amazon EMR.

Tableau Prep allows us to work with big data sets by leveraging sampling. However, it processes data on your local machine and if you want to create, extract, or export data into a CSV using a huge dataset, it can fail due to lack of memory. We learned that Tableau Desktop works with big data by rendering results using a live connection. We don't want to create an extract when working with big data. In the case of Tableau Prep, we can learn our dataset and then use filters to split the dataset and work with part of it.

Note

Here's another solution: we can launch a powerful AWS EC2 instance and install Tableau Prep there, where it will use more resources.

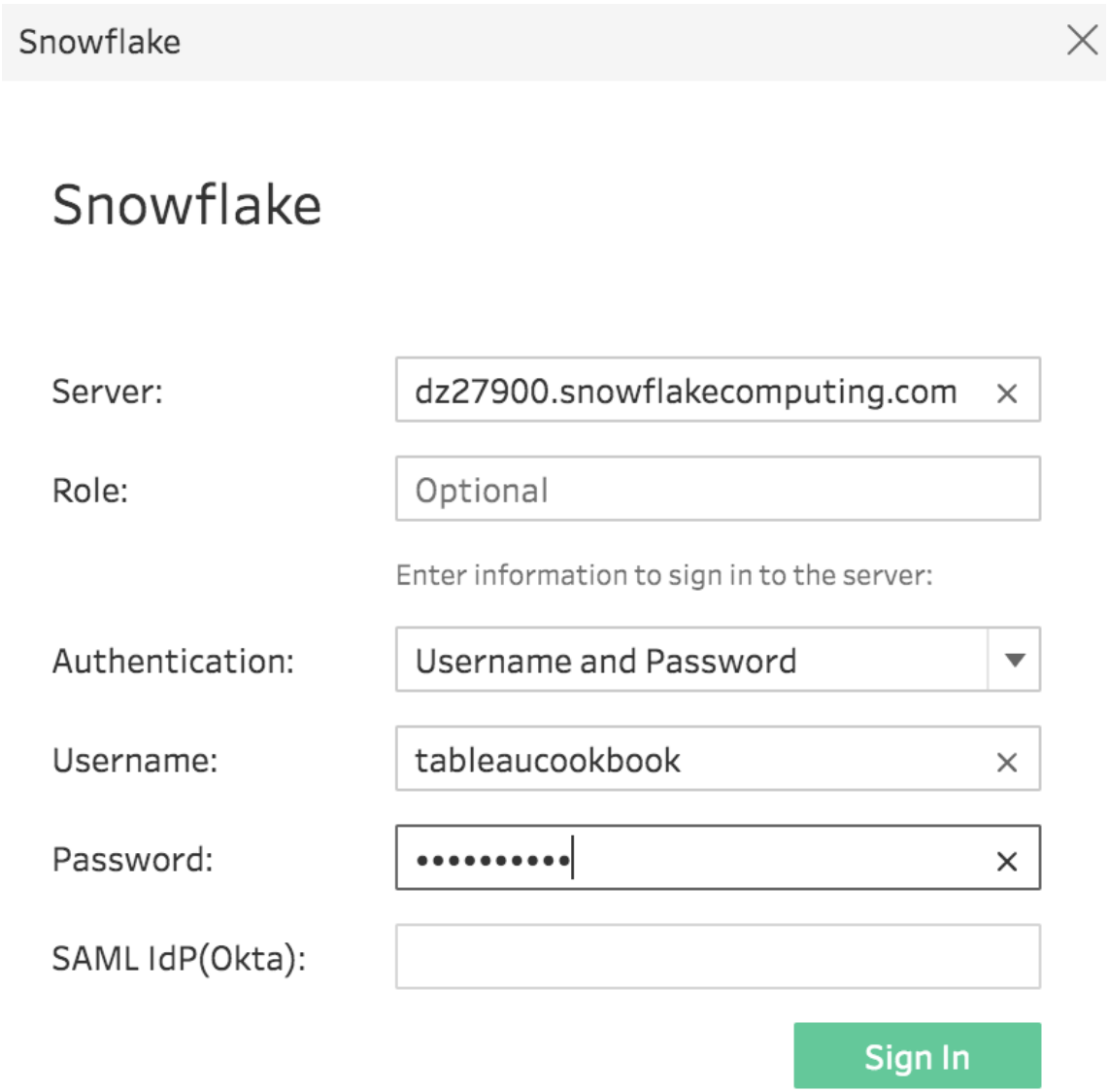
Getting ready

In this recipe, we'll connect our Snowflake cluster and create a flow using Snowflake data in order to calculate metrics by marketing segment.

How to do it...

Let's connect Snowflake and build our flow:

1. Click on `Connections` and choose Snowflake. Fill in the credentials:



2. Choose the `Virtual Warehouse` (computing resource), `Database`, and `Schema` options. It's the same as we did in Tableau Desktop:

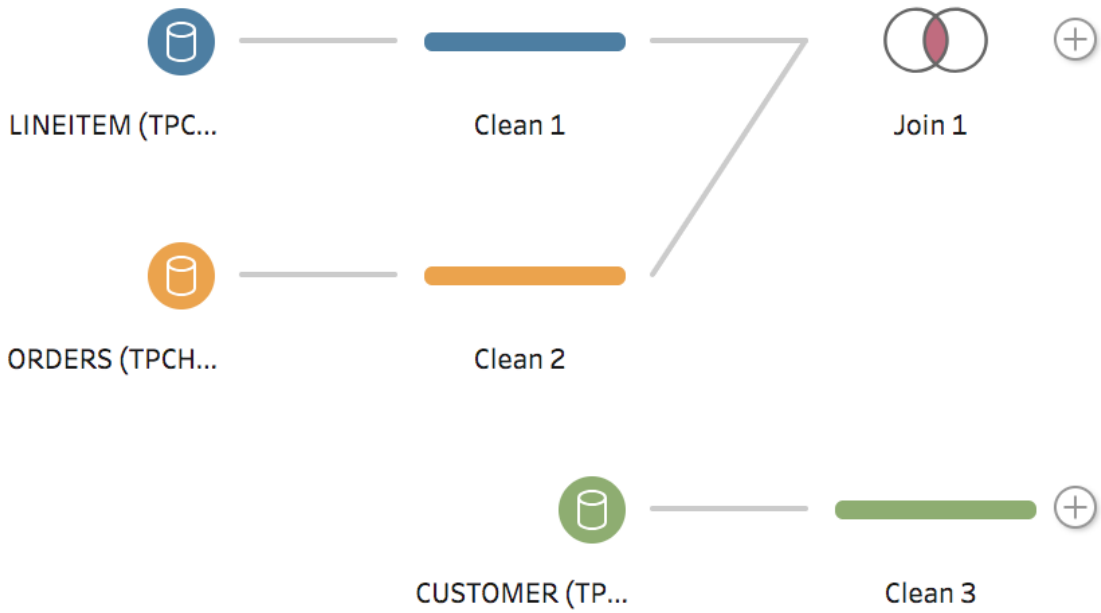
`Warehouse`	`SF_TUTS_WH`
`Database`	`SNOWFLAKE_SAMPLE_DATA`
`Schema`	`TPCH_SF1`

You might choose a different schema with a bigger dataset, such as `TPCH_SF10`, `TPCH_SF100`, or `TPCH_SF1000`.

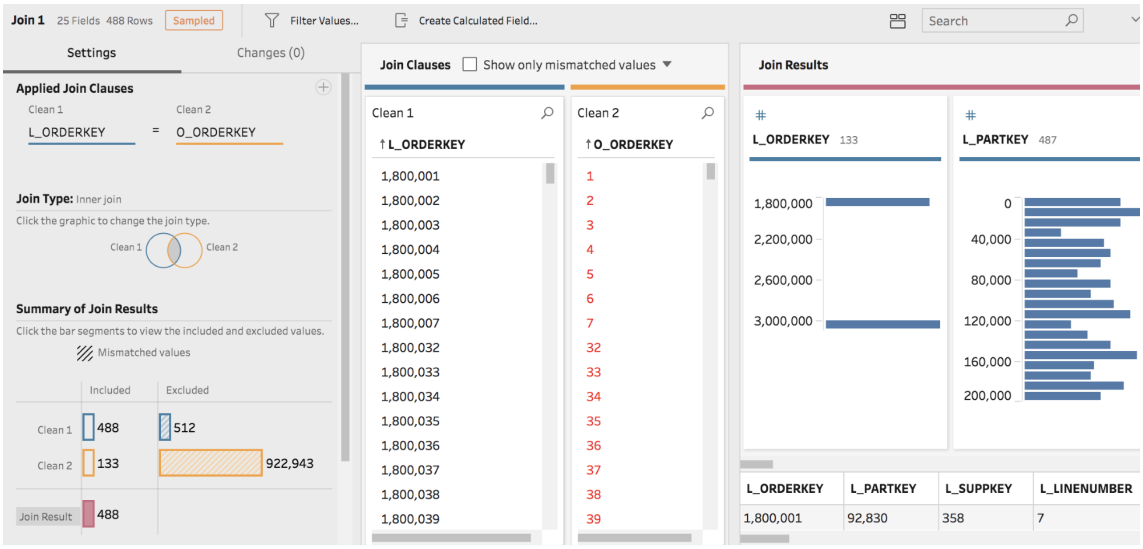
3. Drag and drop tables onto the canvas. You should already know the differences between Desktop and Prep. Let's drag and drop the following tables:
    - **`LINEITEM`**
    - **`ORDERS`**
    - **`CUSTOMER`**

Note

We need to join them all together. In Tableau Prep, we can join only two streams at once. We should join the `LINEITEM` and `ORDERS` tables together. But before joining the dataset, we can learn more about tables and their data using the **`Clean`** step. If we don't want to change any data in the tables, the **`Clean`** step is optional. Anyway, it's a good practice to use the **`Clean`** step after every operation. This is how it should look now:



4. Click on **`Join 1`** and you'll see the **`Profile`** pane. It has nice visualization of the **`JOIN`** statement:



In the `O_ORDERKEY` data card, you might see lots of red values. It means these values aren't joined. Due to the sampling of `LINEITEM`, we don't have the full dataset here. If you want to check the full dataset, you should adjust the `LINEITEM` source table component and eliminate sampling. We will do this at the end of the flow.

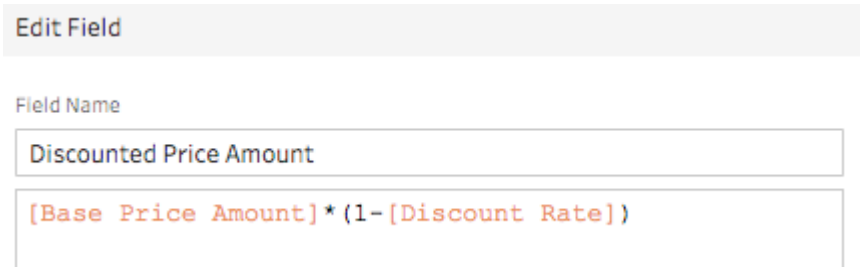
Note

There's another trick to working with big datasets. You might apply filters for the `LINEITEM` data source and filter one or multiple. For example, you might add the `[L_ORDERKEY]=1]` filter. It will bring us only three rows and add transparency.

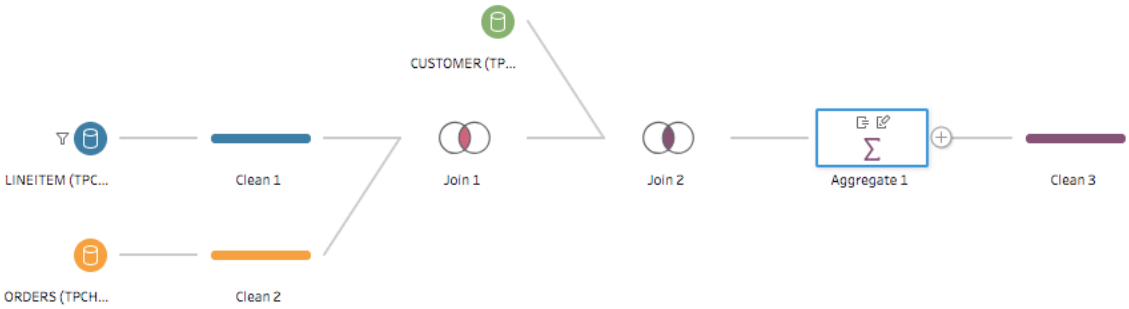
5. Let's add the `CUSTOMER` table. Drag and drop the `CUSTOMER` object to `JOIN 1` and choose the `JOIN` operation. It will create `JOIN 2`. Because we filtered our `LINEITEM` table, we can check that we have only one customer in the flow.
6. Let's create some metrics. Add the `Add Aggregate` step by clicking the `+` sign. At the `Profile` pane, we have `Grouped Fields` and `Aggregated Fields`.
7. Drop `C_MKTSEGMENT` into `Grouped Fields` and then drop the metric fields into `Aggregated Fields`; in addition, we will rename them:

Original name	New name	Function
`L_Quantity`	`Quantity`	`SUM`
`L_EXTENDEDPRICE`	`Base Price Amount`	`SUM`
`L_DISCOUNT`	`Discount Rate`	`SUM`
`L_TAX`	`Tax Rate`	`Sum`
`C_MKTSEGMENT`	`Marketing Segment`	`n/a`

8. Let's calculate some additional metrics, such as amount with discount and tax. We can create calculated fields at the `Add Aggregate` step:

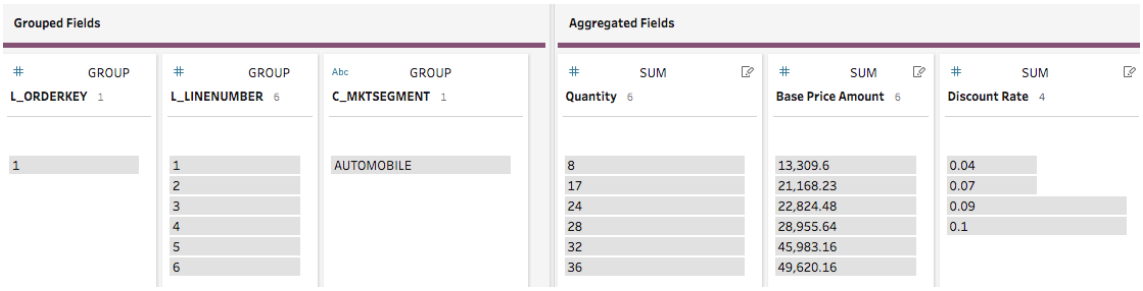


And then we can test the result of this job by adding the **`Clean`** step:



If we check `Discounted Price Amount` at the **`Clean 3`** step, we'll find that it's wrong because we broke the level of aggregation. Based on our dataset, we have a discount rate on the line-item level. This means we should calculate the **`Discounted Price`** on the **`Item`** level and then aggregate. We aggregated the discount on the **`Marketing Segment`** level (on the customer level because we have only one customer).

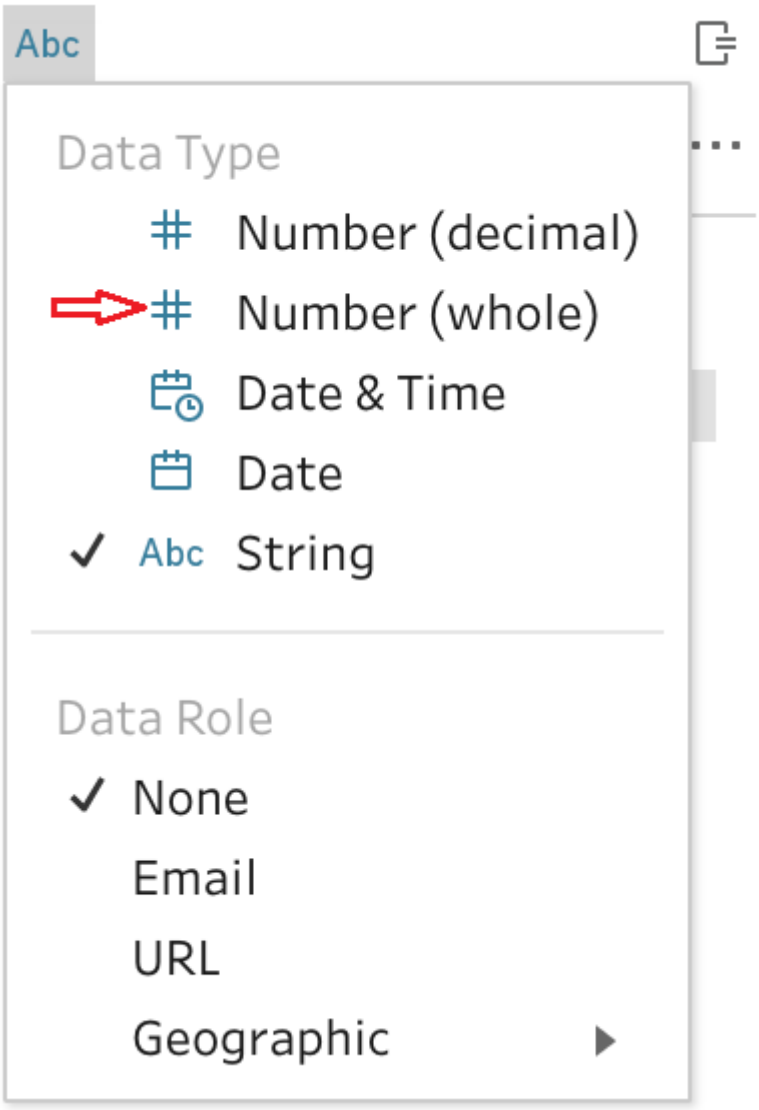
In order to fix this, we need multiple aggregate steps. We should change the existing one, and group by **`Line Item`**, **`Order Key`**, and **`Marketing Segment`**. Then we'll add one more aggregate step where we'll aggregate only on **`Marketing Segment`**:



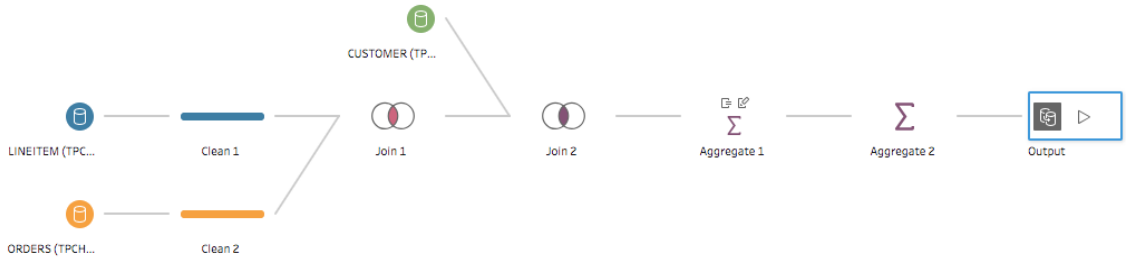
You can see that we have six line items and four different discount rates.

Add one more aggregation step and aggregate on **`Marketing Segment`**. As a result, we'll get the right discounted price.

9. Adjust the data type by changing it from **`Number (decimal)`** to **`Number (whole)`**. In order to do this, just click on data type symbol at data pane:

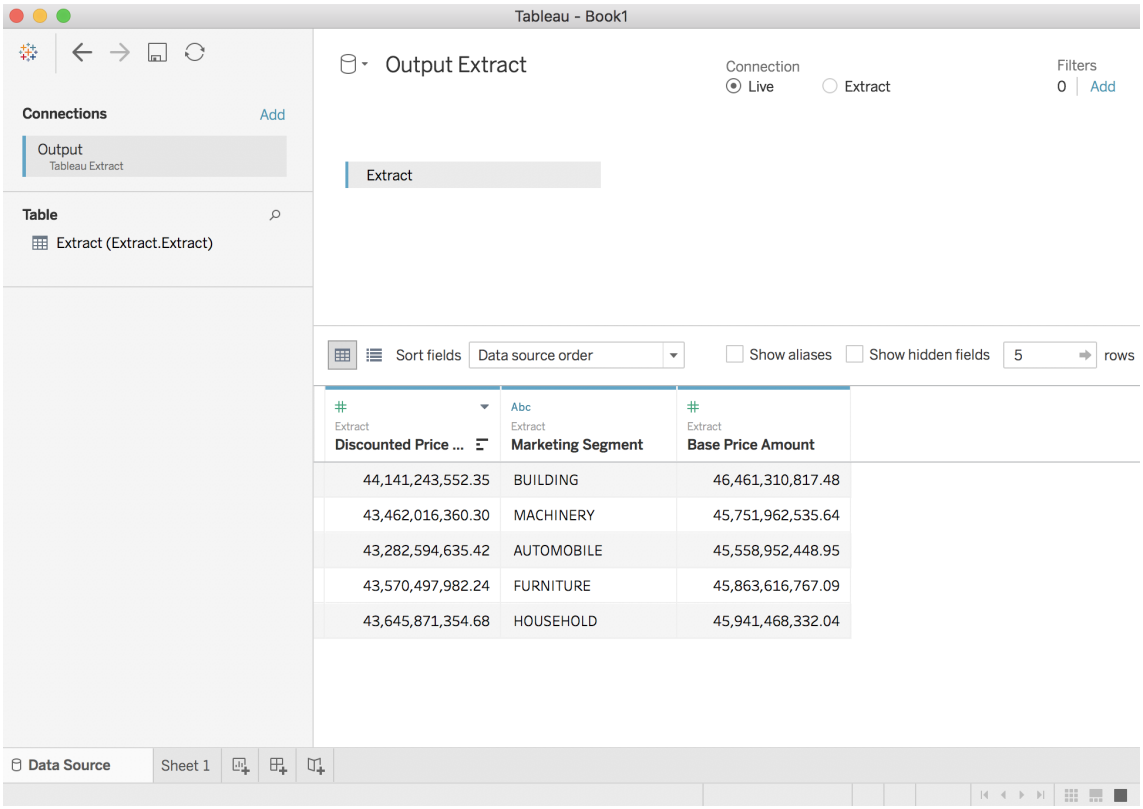


10. Create the final step by adding `Output`. Before running this, drop the filter and eliminate the sampling:



It took 36 seconds to run.

11. Test the result by opening the data source in Tableau Desktop. The data source was saved in My `Documents` | `Tableau Prep` | `Data Sources` | `Output.tde`. Open this with Tableau Desktop:



How it works...

We can go to the Snowflake console and see how Tableau Prep works under the hood:

✓	27b9f9fb-a...	SELECT "b15f6e92-e6c9-4137-a0a7-9c564c"."L_DISCOUNT" AS "L_DISCOUNT", "b15f6e9...	TABLEAUCOOKBOOK
✓	7aef9df4-f5...	SELECT "6d285608-20ca-4f84-b594-65f22f"."O_CUSTKEY" AS "O_CUSTKEY", "6d285608...	TABLEAUCOOKBOOK
✓	d584dff0-fd...	SELECT "edcd0042-63f1-4afb-b973-ee02b4"."L_ORDERKEY" AS "L_ORDERKEY", "edcd0...	TABLEAUCOOKBOOK
✓	edd60452-...	SELECT "287ec79d-633b-4e20-903d-22ef49"."C_CUSTKEY" AS "C_CUSTKEY", "287ec79...	TABLEAUCOOKBOOK

You can see that Tableau Prep ran multiple queries. The first three queries retrieved data from the tables. In the last one, we retrieved data for the first aggregate step. In other words, Tableau Prep is only partly using Snowflake to get the initial data. All other transformations are happening inside of Tableau Prep.

There's more...

You might think of this flow as the logic for a couple of metrics. With Tableau Prep, you can build as many streams as you want, then union them and create the Tableau Extract. For example, we could add a branch in the middle of my flow and start to create new metrics with different grouping options, and then Union with my initial flow and write to the extract. This gives us performance benefits because our dashboards work faster without complex calculations and filters. Moreover, it visualizes the flow and allows end users to quickly understand the logic or apply the changes. There is a link with more information about Tableau Prep steps: https://onlinehelp.tableau.com/current/prep/en-us/prep_clean.htm

See also...

With Tableau Prep 2019.1 was released new feature --- Tableau Prep Conductor. Prep Conductor is that server integration, and it unleashes the full potential of Tableau Prep to operationalize your data prep experience. With Prep

Conductor, you can schedule your flows to run, when, where, and how you want to. You can choose which outputs to schedule independently of one another. You can read more about this feature here https://www.tableau.com/about/blog/2018/11/keep-your-data-fresh-tableau-prep-conductor-now-beta-97369.