Data preparation and data connections in Tableau

Key2Market · Follow

10 min read · Jun 4, 2018

Before you begin your analysis in Tableau, you must connect to your data and then set up the data source. But to successfully upload data, it must meet certain criteria. Also, data can be placed in different types of sources, connection with which has its own characteristics. In addition, Tableau has a wide range of data processing functions, which will help to reduce clutter and make the analysis more efficient and faster. This article is devoted to the aforementioned problems.

Data sources and data connections

Tableau gives you the option to connect to an Excel/csv, Text, Access or other file located on your PC or Mac, or the option to connect to a server. The Excel, Text, Access, and Statistical files will be local databases, such as files located on your own computer, while a server will allow you to connect to an online data source.

Let's start with an easy type of connection — **to a file**. For example, you want to connect to Excel file. In the block "Connect" => "To a File" click "Microsoft Excel", in the opened dialog box select the Excel workbook you want to connect to. But there is another way. You can just open your Excel workbook, select and copy data, then on Tableau panel click "File" => "Paste", and the data will be loaded into a new workbook.

Also, you can connect **to a server**. This means that data you want to connect to locates on the server (not your computer). Let's consider a special type of connection — "Tableau Server". In order to use this type of connection user need to have an access to Tableau server and any data source should be available there. For this, start Tableau Desktop, and under "Connect", select "Tableau Server", enter the name of the server and then select "Connect". Next you need to enter login and password and then choose a data source you need from available ones.
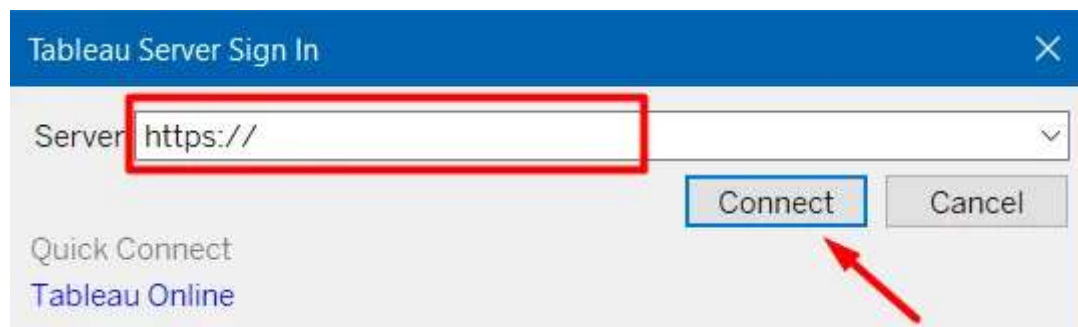

Tableau server sign in window

Also, you can connect to any other servers available under "Connect to a server" type. For example, below is the screenshot of connection window to Amazon Redshift, but it will be pretty much the same for any other data base.

Amazon Redshift

Server: [   ] Port: [   ]

Database: [   ]

Enter information to sign in to the database:

Username: [   ]

Password: [   ]

☐ Require SSL

Initial SQL...   Sign In

Fill in five required fields with your credentials and you'll be able to use DB data for your analysis

Preparing data before connection

Before connecting to your data, you need to make sure that the data is correctly prepared. For Tableau, there are three steps to format your data correctly. Let's consider an example of connection to Excel file.

First, make sure that your data begins in column A1 and that the first row of data only includes column headers. Second, make sure each column of data contains the same type of data throughout the column and each row of data is one data item. Finally, remove any column or row totals and introductory text. Data should not be organized in a crosstab, aggregated, or contains other extraneous information. Make sure that

there are no blank rows in your data. Once your data is correctly formatted, you are ready to import the data into Tableau.

	A	B	C	D	E	F	G	H
1	Sales Record Number	Buyer City	Item Title	Quantity	Sale Price	Sale Date	Feedback Left	Custom Label
2	12000	Lachine	Cressi Leonardo Dive Computer Watch	1	$199.95	Apr-01-18	No	211157-35-A
3	12001	San Antonio	Cressi Leonardo Dive Computer Watch	1	$199.95	Apr-01-18	No	211157-35-A
4	12002	TEMECULA	Cressi Leonardo Dive Computer Watch	1	$199.95	Apr-01-18	No	211157-35-A
5	12003	Skiatook	Cressi Leonardo Dive Computer Watch	2	$199.95	Apr-01-18	No	211157-35-A
6	12004	Tucson	Cressi Leonardo Dive Computer Watch	1	$199.95	Apr-01-18	No	211157-35-A
7	12005	West Berlin	EEZYCUT Harness Pouch	1	$13.95	Apr-01-18	Yes	216552-A
8	12006	El Paso	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	Yes	216552-A
9	12007	hanover	EEZYCUT Harness Pouch	4	$13.95	Apr-02-18	No	216552-A
10	12008	San Diego	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
11	12009	Norcross	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
12	12010	Wilmington	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
13	12011	Goodyear	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
14	12012	Omaha	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
15	12013	jenks	EEZYCUT Harness Pouch	2	$13.95	Apr-02-18	No	216552-A
16	12014	Miami	EEZYCUT Harness Pouch	1	$13.95	Apr-02-18	No	216552-A
17	12015	Honolulu	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-02-18	Yes	213949-A
18	12016	Channelview	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
19	12019	huntsville	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
20	12020	Miami	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
21	12021	Kensington	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
22	12022	KANNAPOLIS	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	Yes	213949-A
23	12023	Franklin	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
24	12024	Mc Gregor	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
25	12025	Concord	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
26	12026	Brownsville	XS Scuba Neoprene Knife Wrap	1	$16.00	Apr-05-18	No	213949-A
27	12027	McHenry	Silicone Grease - 2oz	1	$7.95	Apr-05-18	Yes	213591-A
28	12028	Chepachet	Silicone Grease - 2oz	1	$7.95	Apr-05-18	Yes	213591-A
29	12029	Aurora	Silicone Grease - 2oz	1	$7.95	Apr-05-18	No	213591-A

A good example of spreadsheet which may serve a data source for Tableau

	A	B	C	D	E	F	G	H	I
1		Sales Record Numbe	City	Item Title	Quantity	Sale Price	Sale Date	Feedback Left	Custom Label
2		12000	Lachine	Leonardo Dive Computer	1	199.95	Apr-01-18	No	211157-35-A
3		12001	San Antonio	Leonardo Dive Computer	1	199.95	Apr-01-18	No	211157-35-A
4		12002	TEMECULA	Leonardo Dive Computer	1	199.95	Apr-01-18		211157-35-A
5		12003	Skiatook	Leonardo Dive Computer	2	199.95	Apr-01-18	No	211157-35-A
6		12004	Tucson	Leonardo Dive Computer	1	199.95	Apr-01-18	No	211157-35-A
7		12005	West Berlin	EEZYCUT Harness Pouch	1	13.95	Apr-01-18	Yes	216552-A
8		12006	El Paso	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	Yes	216552-A
9		12007	hanover	EEZYCUT Harness Pouch	4	13.95	Apr-02-18	No	216552-A
10		12008	San Diego	EEZYCUT Harness Pouch	1	13.95	Apr-02-18		216552-A
11		12009	Norcross	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	No	216552-A
12		12010	Wilmington	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	No	216552-A
13		12011	Goodyear	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	No	216552-A
14		12012	Omaha	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	No	216552-A
15		12013	jenks	EEZYCUT Harness Pouch	2	13.95	Apr-02-18	No	216552-A
16		12014	Miami	EEZYCUT Harness Pouch	1	13.95	Apr-02-18	No	216552-A
17		12015	Honolulu	Scuba Neoprene Knife Wr	1	16	Apr-02-18	Yes	213949-A
18		12016	Channelview	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
19		12019	huntsville	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
20		12020	Miami	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
21		12021	Kensington	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
22		12022	KANNAPOLIS	Scuba Neoprene Knife Wr	1	16	Apr-05-18	Yes	213949-A
23		12023	Franklin	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
24		12024	Mc Gregor	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
25		12025	Concord	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
26		12026	Brownsville	Scuba Neoprene Knife Wr	1	16	Apr-05-18	No	213949-A
27		12027	McHenry	Silicone Grease - 2oz	1	2+5.50	Apr-05-18	Yes	213591-A
28	Total	27	-	-	32	-	-	-	-

A bad example: filters are on, data starts from B2 cell, column totals aren't deleted etc.

The distinctive feature of the connection to data base server is that the data in the data source can be aggregated or mixed in the way you need inside data base (if possible) to ensure higher efficiency of reports.

Take another look at the table above ('right' example) and suppose it is in a data base. For example, you want to analyze only sales by day, it makes sense to **create a view**: "*CREATE VIEW sales_by_day AS sales_date, SUM(sales_price) FROM [table name] GROUP BY sales_data*". With a small amount of data, this does not play a significant role, but if your table has hundreds of thousands or millions of rows, this approach will significantly accelerate calculations and reduce operating time of Tableau workbook. Finally, when you connect to the database, Tableau will download the aggregated data on the above criteria.

Another way to transform a data in data is to create **joins and unions** inside data base. More on these features (but in Tableau) in next section.

Data preprocessing inside of Tableau (on the "Data" tab)
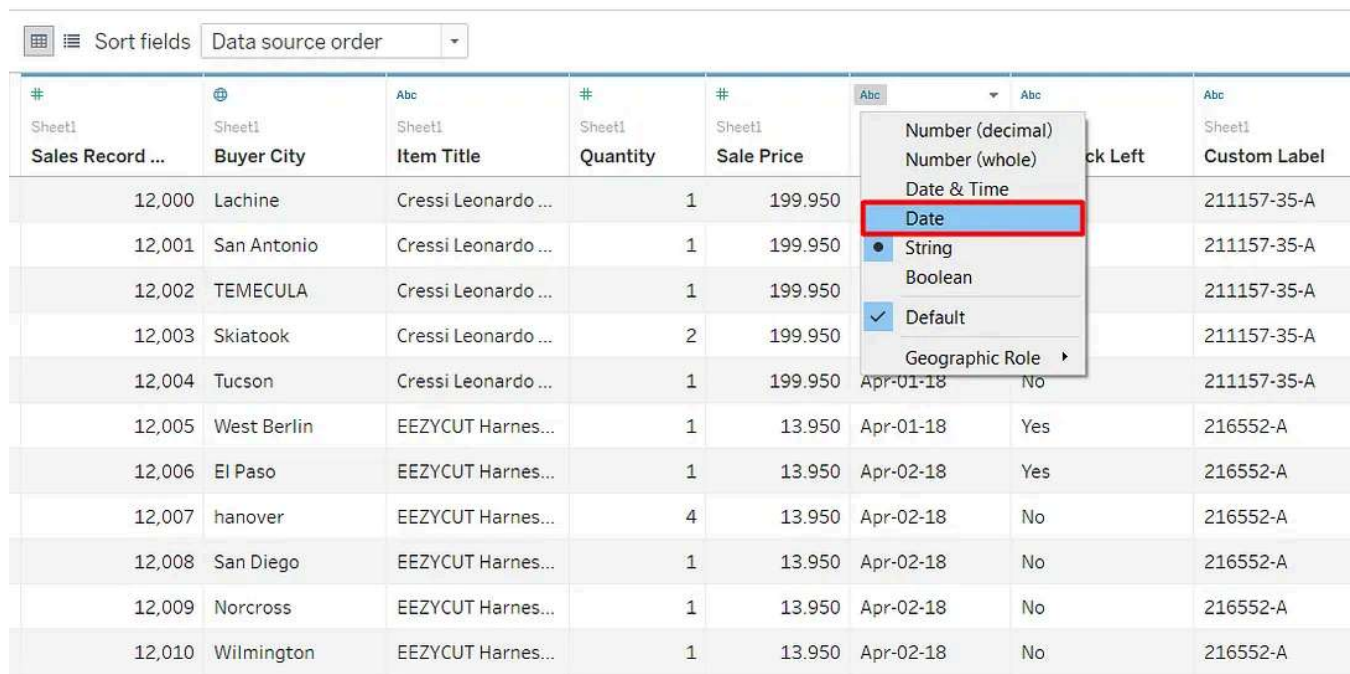
1. DATA TYPES

Since each column has a header name and the data type is automatically identified by Tableau through the use of icons. An ABC icon represents string data, for example, data that includes names. A globe icon represents geographical data, for example, data that includes cities, states, zip codes. Also, you can assign a **geographic role** for those fields if Tableau determined it incorrectly.
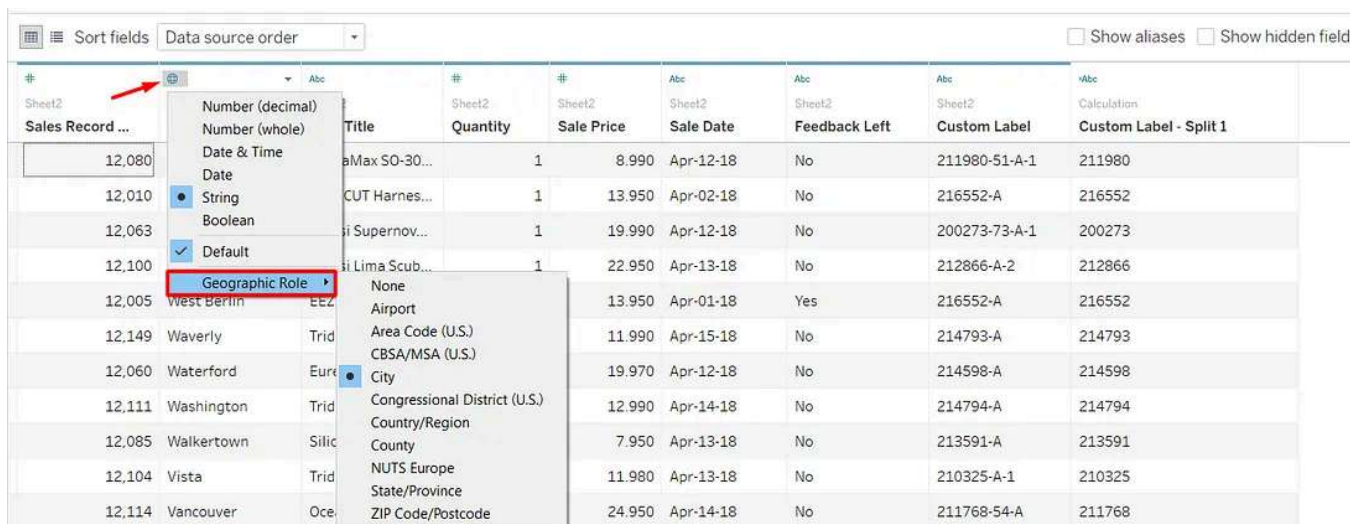
You can change geographic role manually, although usually Tableau is quite smart to determine the right type automatically

A calendar icon represents dates, and the pound icon represents numeric data. If Tableau incorrectly assigned a data type, you can click on the associated icon and

choose the correct data type.

Data type can be changed manually as well
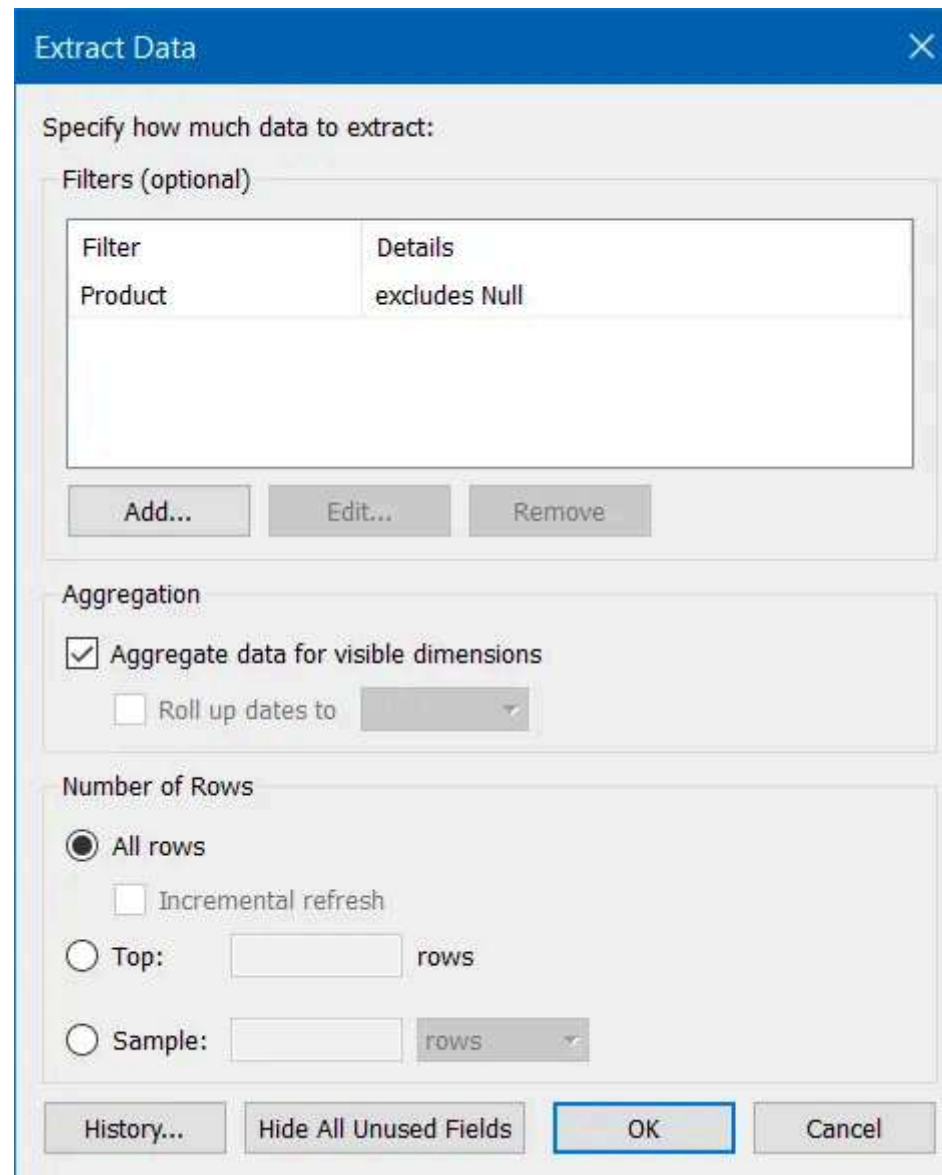
2. CONNECTION METHODS

There are two primary types of connections for Tableau. A live connection and data extract. A **live data connection** means that Tableau sends queries to your underlying database and retrieves up to date data whenever the Tableau dashboard is updated. A live data connection is best and should be used when the live connection between the Tableau dashboard and your data is possible, required by end user and impact on report speed (processing time) is moderate. The other type is a Tableau data **extract connection**. Which means that Tableau sends queries to your extracted or static database and not the underlying data. Any updates to the

underlying database will not update the Tableau dashboard. You will have to refresh the Tableau extract to get the up-to-date data. Extract connection should be used when a live connection to your data is not possible, when the live data connection is too slow, or in such cases where the size of the data is extremely large.
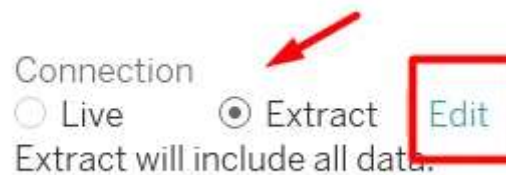
To make the visualization up to date, but at the same time to reduce the time of its loading, it makes sense to use the **updated extracts on a server**. Tableau will use the data from the published extract, which will be updating at a given frequency. It's very simple to do this. After you set up the data source on the Data Source page, in the upper-right corner, select "Extract", and then click "Edit". Click "Add" to define one or more filters to limit how much data gets extracted based on fields and their values. Select "Aggregate data" for visible dimensions to aggregate the measures using their default aggregation. Aggregating the data consolidates rows, can minimize the size of the extract file, and increase performance. You can extract All rows or the Top N rows. When finished, click OK and then the "sheet" tab. Clicking the sheet tab initiates the

creating of the extract. In the subsequent dialog box, select a location to save the extract, give the extract file a name, and then click "Save".

Then you need to connect to the server, log in, and publish the extract by filling out the fields in the following form:

After specifying the necessary data, you need to select the frequency with which the extract will be updated. For example, the extract will be updated every day at 2:00.

But, if necessary, you can update the data manually at any time.

3. WORKING MULTIPLE TABLES

There are a few different types of joins that Tableau will provide. An inner join, left join, right join, or a full outer join. Let's consider each one of these and how or why you would use them. An **inner join** will show the records or rows present in both tables, if there is at least one match between columns. You will always use an inner join when you want to display all of the data that is stored in both of the data sources, using a merging variable. A **left join** returns all of the records or rows present in the left table and matching rows from the right table. You would use a left join when you want to keep all of the data in the left table and only merge in limited data from the table on the right. A **right join** returns all of the records or rows present in the right table and matching rows from the left table. You would use a right join when you want to keep all of the data in the right table and only merge in limited data from the table on the left. A **full outer join** were shown all of the records or rows present in both left and the right table. You will use a full outer join when you wanted display all of the data that is stored within all of the tables If the data joined, chosen by Tableau, is not the join you wanted, you can simply click on the join to open the join dialog box.

You can **union** your data to combine two or more tables by appending values (rows) from one table to another. To union your data in Tableau data source, the tables must come from the same connection. For best results, the tables that you combine using a union must have the same structure. That is, each table must have the same number of fields, and related fields must have matching field names and data types. To create a union, click "New Union" below the list of data connections and add right sheets to union. You can rename union by double-click on "Union" or edit it.

Also, can be used an automatic union, where you can enter the search criteria that you want Tableau to use to find tables to include in the union.

Automatic union form

Optional configurations with data before analysis

1. HIDDEN FIELDS

What should you do when you can't find some field in "Dimensions"-"Measures" blocks, but you know that you data source contains it? For example, you can't find field "Feedback left" in workbook below. The easiest way to check its availability is to check "Show hidden fields" checkbox:

After that you can see a hidden field, and since the field is needed for visualization, it makes sense to unhide it.

2. SPLIT

Another useful function in preparing data for analysis is "**Custom split**". For example, I have a spreadsheet with a column called "Custom Label", since each product has a unique SKU. But each SKU has a root value, which unites the goods of a certain type of one manufacturer, and variable values, which correspond to the different sizes and colors of this product. Let's say we want to split the data by the root SKU. To realize this, we go to tab "Data Source", to click on the down arrow on the field "Custom Label", then "Custom Split", click on that.

We are going to use a separator here — a hyphen, which is exactly how the footnote is described.

And then we're going to split off the first one and then look at that. I have created two variables. There is "Custom Label" which is original. Then there is "Custom Label-Split1"which takes the values from the SKU, and splits it at that hyphen.

3. PIVOT

In the process of preparing information for analysis, you may need a function such as "Pivot", which can pivot your data from crosstab format into columnar format. You

need to set up your data source and then select two or more columns (holding down CTRL), click the drop-down arrow, and then select "Pivot". New columns called "Pivot field names" and "Pivot field values" replace the original columns that you selected to create the pivot.

Also, you can add columns to "Pivot" and delete it at all, after which the table will return to the original view.

4. FILTERS

Tableau gives you the ability to **create filters** on a data source (it locates on the top right area of the Data tab). You can "Add" filter manually on the tab "Data Source", then select a field to filter, and finally specify how the field should be filtered. Created filter can be edited or removed if you need.

Data Science   Tableau   Data Visualization   Data Analysis   Data Processing

Written by Key2Market

50 Followers · 4 Following

We help companies set up best Business Intelligence practices

Responses (1)

What are your thoughts?

Respond

Alex Miner
Mar 5, 2020

Thanks for writing this detailed tutorial. If somebody would like to automate the data preparation process without setting it up manually in Tableau, I can advise checking out the no-code data integration approach…

more

Reply

More from Key2Market

Key2Market

Tableau Performance Optimization

When we work with large amount of data, multiple data sources, dashboards and workbooks, which heavy loaded with individual views and…

Dec 5, 2018   👏 35   💬 1

Key2Market

Sending Apache Airflow Logs to S3

I have spent majority of the day today figuring out a way to make Airflow play nice with AWS S3. Not that I want the two to be best…

Aug 18, 2018   👏 30   💬 1

Key2Market

Custom date periods in Tableau

In this post we will show you how to create custom date periods in Tableau.

Sep 3, 2018   👏 56

Row-Level Security (RLS) in Tableau

Key2Market

Row-Level Security (RLS) in Tableau

Nowadays, many organizations face such problems as leaking of confidential information and time-management issues. If you are working in…

Feb 1, 2019   👏 6   💬 1

See all from Key2Market

Recommended from Medium

An ETL pipeline using Pyspark for a Bank

Pushpendra Sharma

An ETL pipeline using Pyspark for a Bank

In the previous article, we performed data wrangling and exploration using Databricks notebooks. While notebooks are great for exploring…

Dec 5, 2024 👏 3

Tableau interview questions for getting or keeping your job. Knowledge is your trump card.

👤 Sharon Fine Point Writing

Tableau interview questions for getting or keeping your job. Knowledge is your trump card.

When it comes to working with Tableau, I quickly learned it's not enough to simply create great visualizations. You must show that you have…
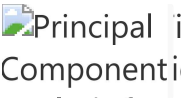
✨ Sep 28, 2024

Lists

	Predictive Modeling w/ Python
	20 stories · 1832 saves

	ChatGPT prompts
	51 stories · 2575 saves

Principal Component	Practical Guides to Machine Learning
	10 stories · 2207 saves

	Coding & Development
	11 stories · 1009 saves

Tableau dashboard featuring Udemy data.

 In Pipeline: Your Data Engineering Resource by Zach Quinn

Creating The Dashboard That Got Me A Data Analyst Job Offer

A walkthrough of the Udemy dashboard that got me a job offer from one of the biggest names in academic publishing.

✦ Dec 5, 2022   👋 2.1K   💬 42

When to Use COUNT(*) vs COUNT(1) in SQL Queries

 Vijay Gadhave

When to Use COUNT(*) vs COUNT(1) in SQL Queries

Note: If you're not a medium member, CLICK HERE

✦ Jan 14   👋 350   💬 25

Cover Image

 J.

Building a Data Pipeline with Python: A Step-by-Step Guide to ETL Processing

✦ Nov 4, 2024   👋 20

The 15 Principles of Data Visualization

In CodeX by Andrew W. Pearson

The 15 Principles of Data Visualization

"Having all the information in the world at our fingertips doesn't make it easier to communicate: it makes it harder," says Cole Nussbaumer…

Aug 25, 2023   513   3

See more recommendations