**Lab 11 - Data Exploration with Dataiku**

**Part 1 - Dataiku - Installation**

There are several methods of installation, click in the table of contents to get to the installation method you want. Please note that there is no instruction for Windows user since DataIku is not supported on Windows. Therefore, Window users should check out the installation section **2. Install on Google Cloud, 4. GNU/Linux Installation or 5. Install Using Virtualbox on VMWare.**

Out of all methods, Google Cloud method is the easiest. I highly recommend that method whether you are a Mac or Windows user.

Table of Contents

[1.Installation Instructions from DataIku 2](#_Toc54877292)

[2. Install on Google Cloud 3](#_Toc54877293)

[3. Install on Mac 6](#_Toc54877294)

[4. GNU/Linux Installation 7](#_Toc54877295)

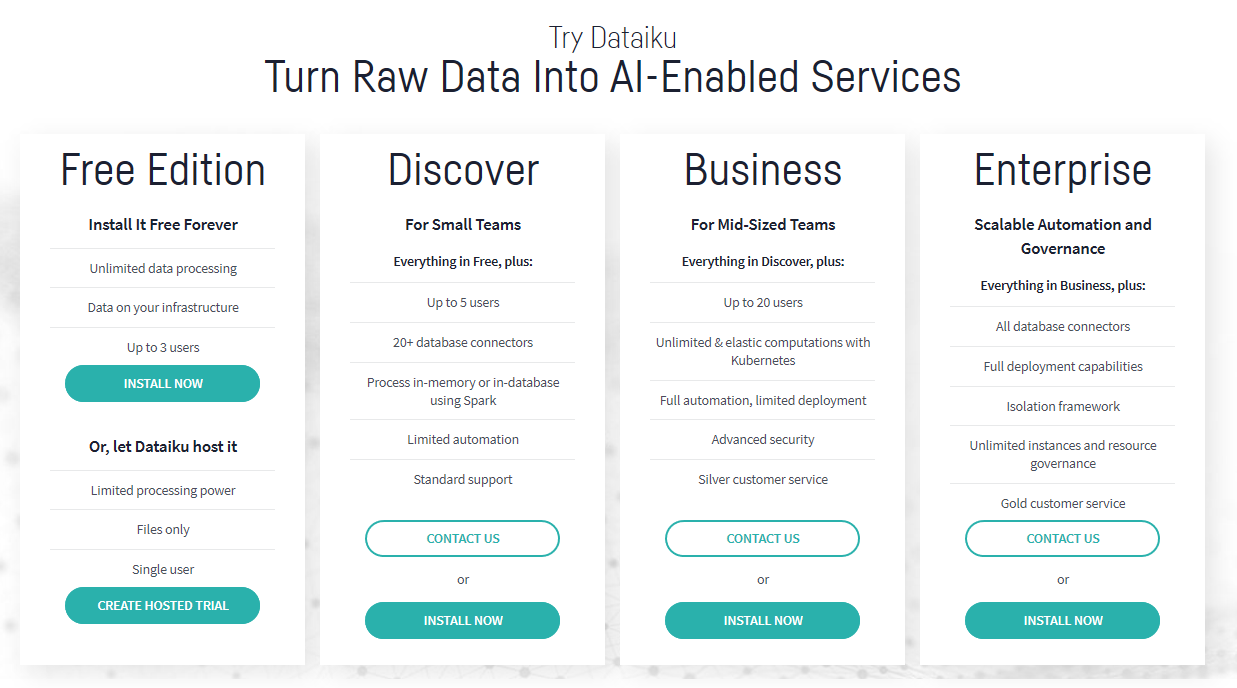
[5. Install Using Virtualbox Or VMWare 8](#_Toc54877296)

[6. Hosted Online Trial by DataIku 10](#_Toc54877297)

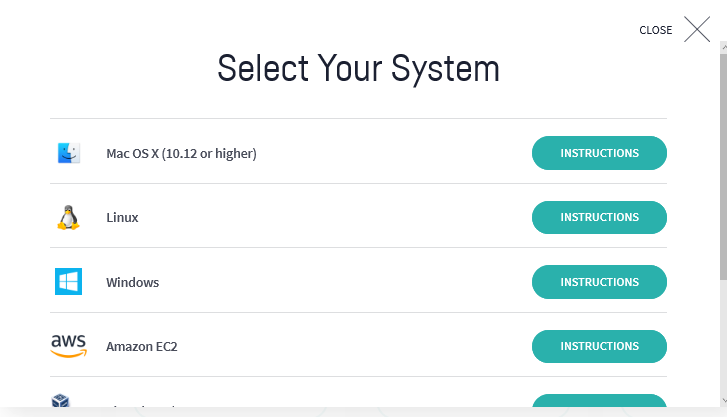
# 1.Installation Instructions from Dataiku

Go to <https://www.dataiku.com/product/get-started/> this page to download dataiku.

Chose the free edition.



Then select your system.



The instructions page DataIku take you to will be similar to the materials present here.

[AWS Instruction Page](https://www.dataiku.com/product/get-started/aws/)

[Mac OS X (10.12 or higher)](https://www.dataiku.com/product/get-started/mac/)

[Linux](https://www.dataiku.com/product/get-started/linux/)

[Virtual Box/ VMWare](https://www.dataiku.com/product/get-started/virtualbox/)

[Google Cloud](https://www.dataiku.com/product/get-started/install-on-google/)

[Microsoft Azure](https://www.dataiku.com/product/get-started/azure/)

If you have Google Cloud or Amazon EC2 account, use this platform to open dataiku is the easiest way. Every computer can open dataiku with google cloud, so in this tutorial, it will show how to use dataiku in google cloud.

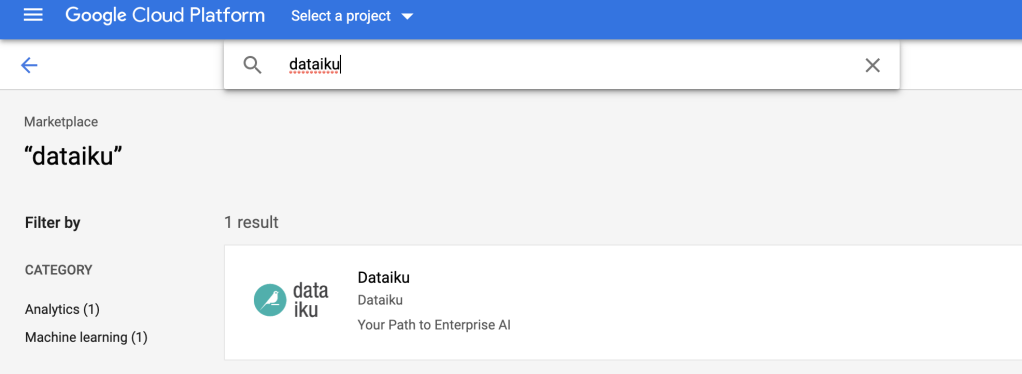
# 2. Install on Google Cloud

First go to <https://cloud.google.com/> and click go to console.

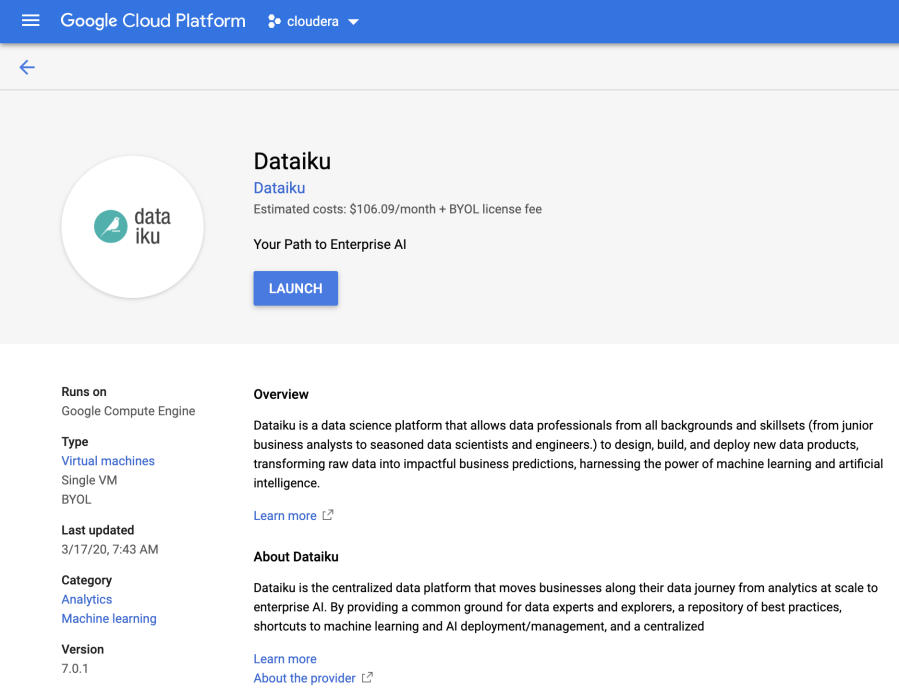
Each account can have $300 free credit trial, you’d better apply for the $300 credit free trial before start. If you once have an account and the free trail has already ended, you can apply another account with a new email address to get another $300 credit.

Step 1: Start the instance

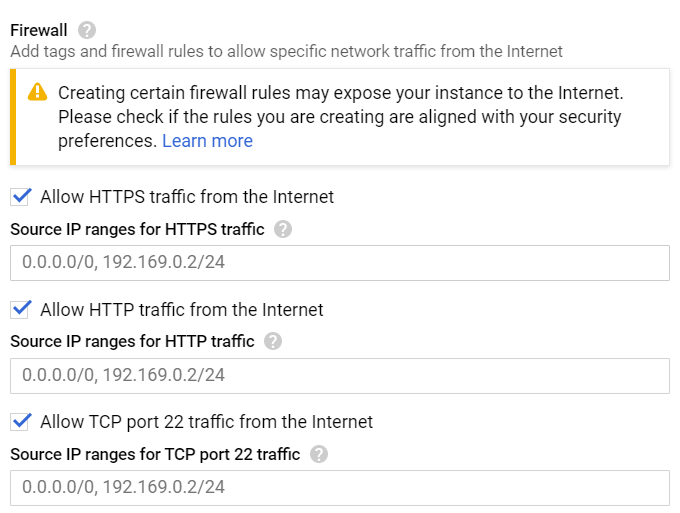
Open the [Dataiku DSS](https://console.cloud.google.com/marketplace/details/dataiku-public/dataiku) page on Google Cloud marketplace



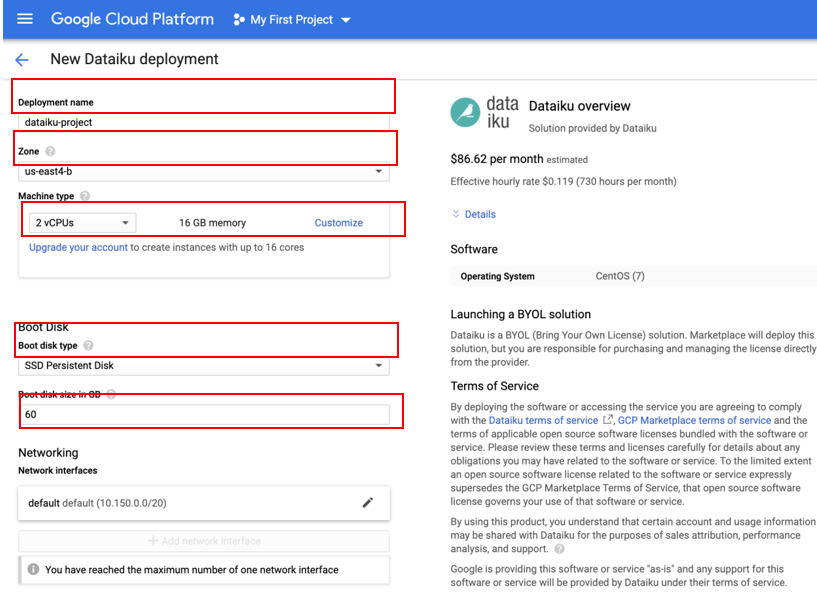
* Click **“Launch on compute engine”**. You will be invited to select the google projects you want to use.



* Select one and click on **Open**
* Enter a name for your Virtual machine in the **“Deployment name”** section and select your region (if not the default) and disk size (at least 50 GB). For example, I selected us-east1-b because I am in NY, therefore the eastern region.
* Make sure you have firewall rules to access the application. Usually it is pre-determined as follow and you do not have to change anything:

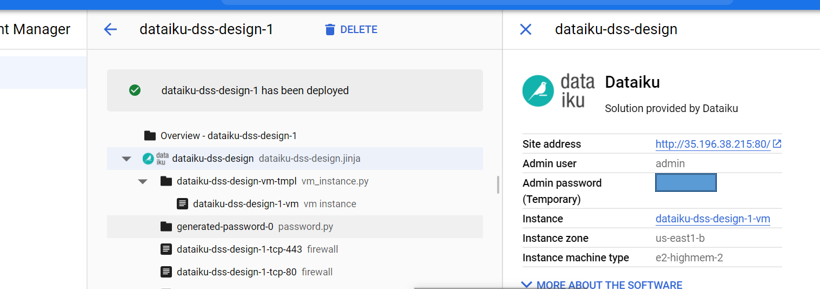


* A machine with 2 vCPUs, 16 GB RAM should be enough for a first test.
* Review the form and then click on **Deploy**.  
  (You may have one warning about the disk size being greater but it’s expected)



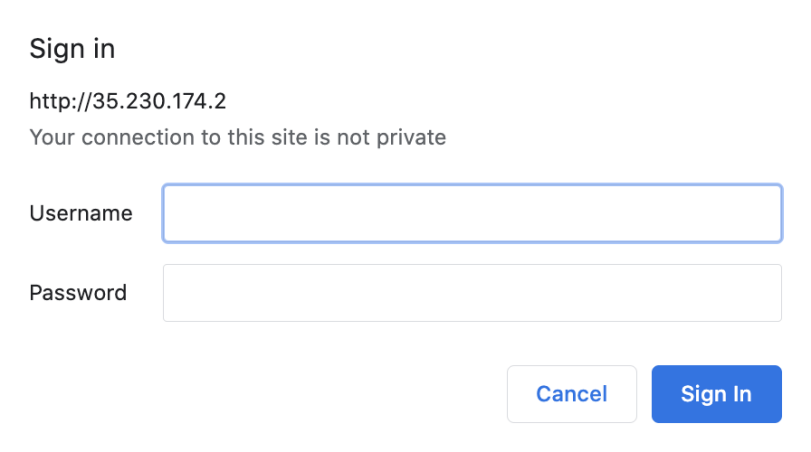
**Step 2: Connect to DSS**

Once deployed, in the Dataiku-Dss-Design section you might see a Site address, **Admin user and Admin password(remember this, you will have to enter them later)**. They are dedicated to the initial authentication on the web UI

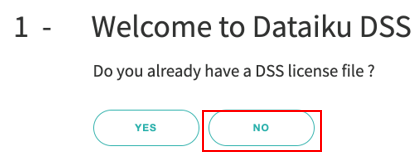


Click on the link in the **Site address** section. Then enter the username and password.

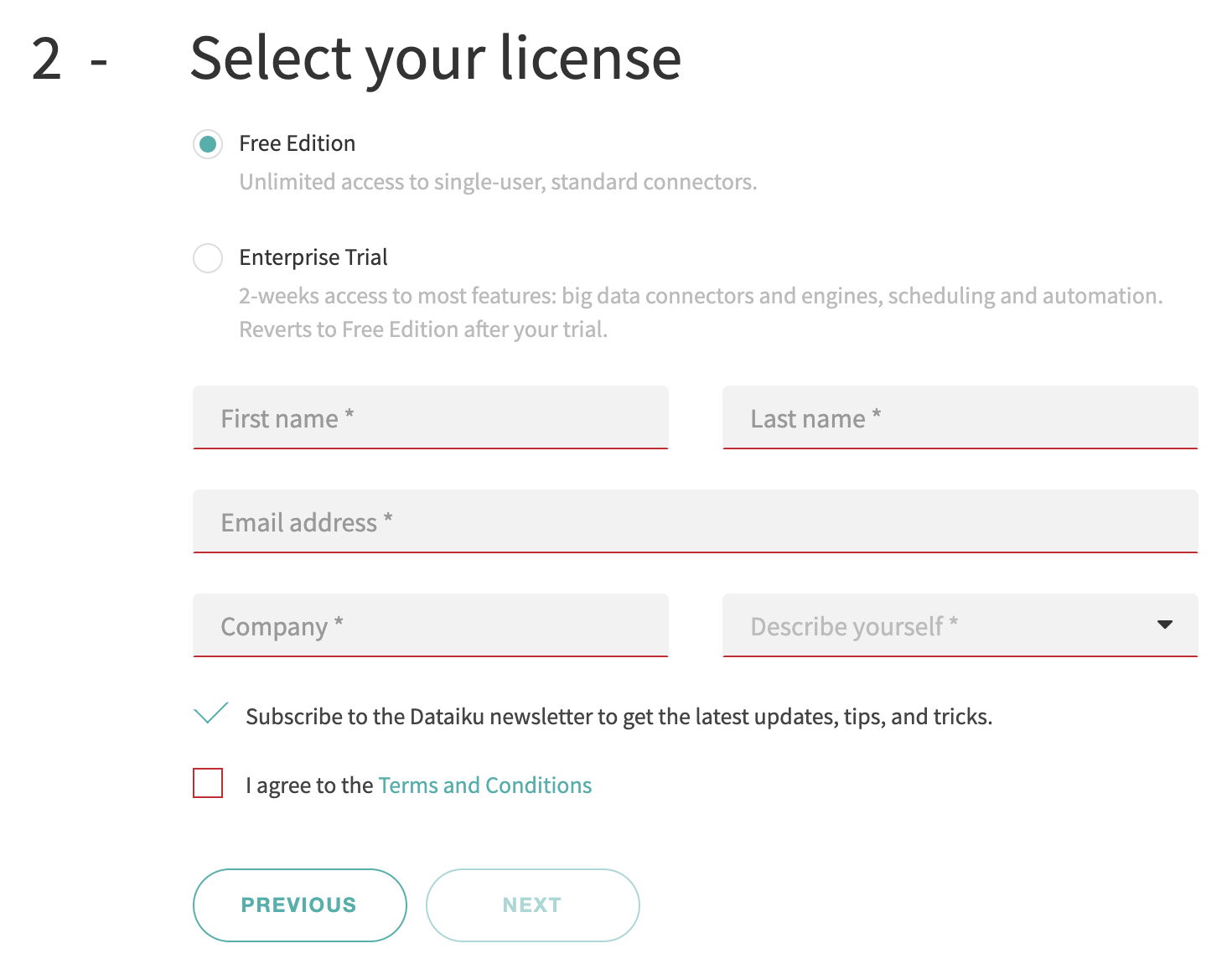
Please note that this is a one-time password only, and after logging in and registering DataIku will give you a different password.



Then you will see this welcome page. Click No to create you Dataiku DSS license.



Then choose the free edition and fill your information.



Then you will have your new login and password

It will be admin for both username and password, unless otherwise stated.

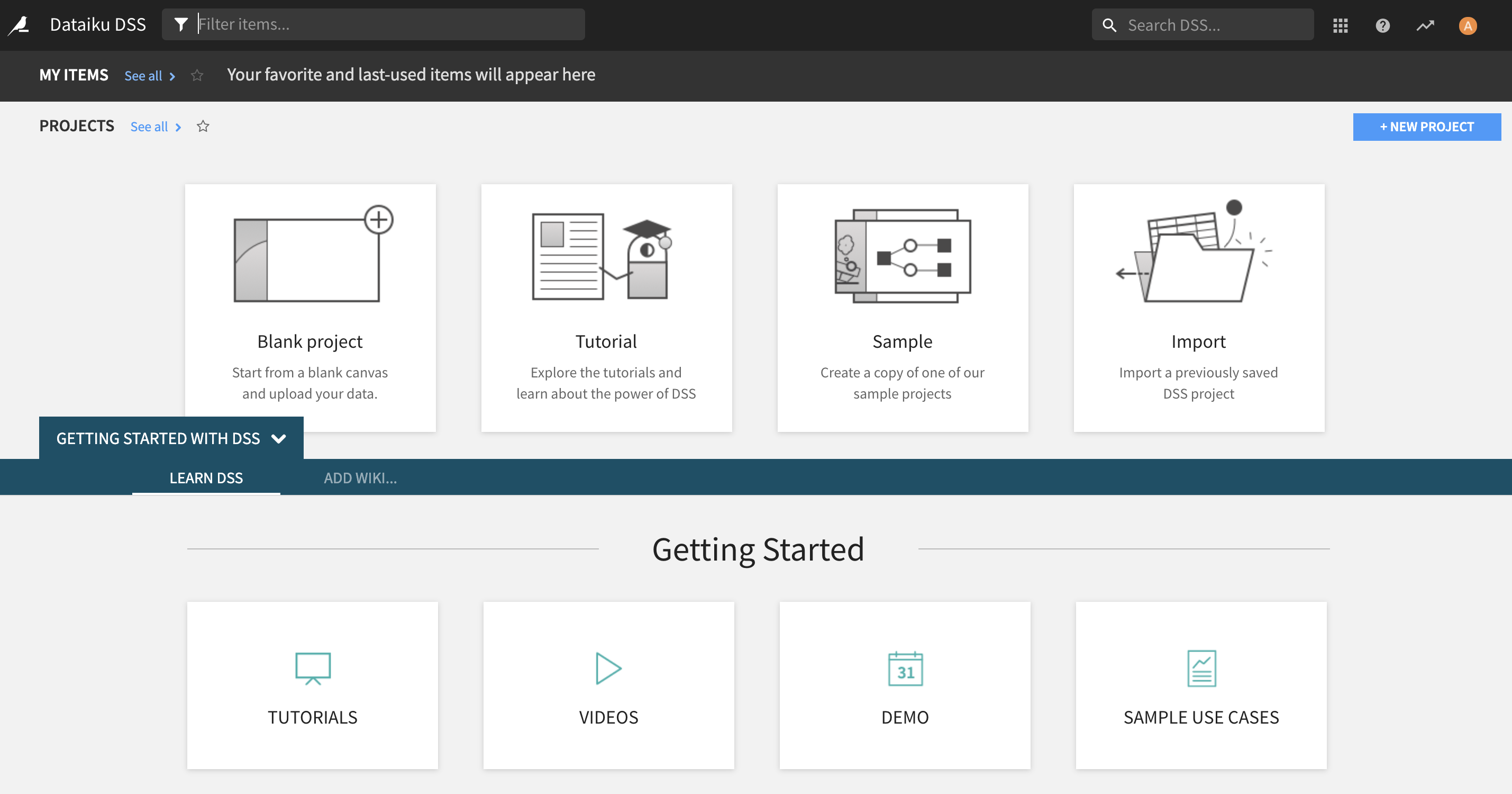
Remember this new log in and password and use it to log in and get started!

Note that only Chrome and Firefox are supported.

For additional information, technical details, how-to, or any issue, please see:

* [Google Cloud installation](https://doc.dataiku.com/dss/latest/installation/other/gcp.html) documentation
* [Community answers](https://answers.dataiku.com/)

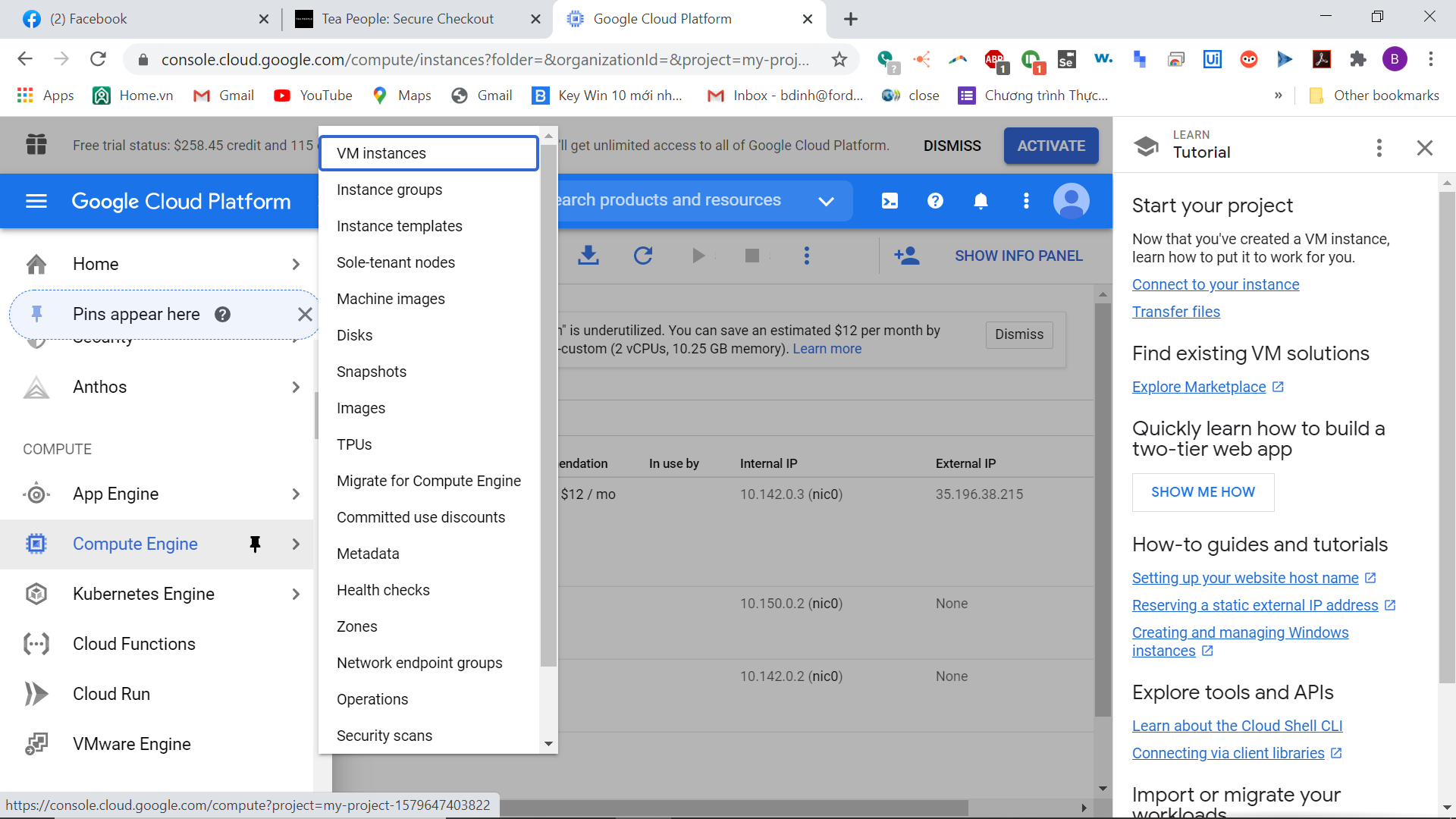
When you see this page, it means you have created your account and opened it successfully.



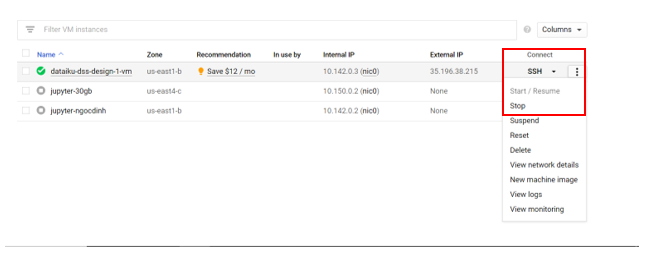
## Logging out of Google Cloud’s DataIku

When you are done, remember to log out.

You also need to shut off the instance in your Google Cloud Console. To do so, open the  on Google Cloud and scroll down to Compute Engine, and select VM Instances

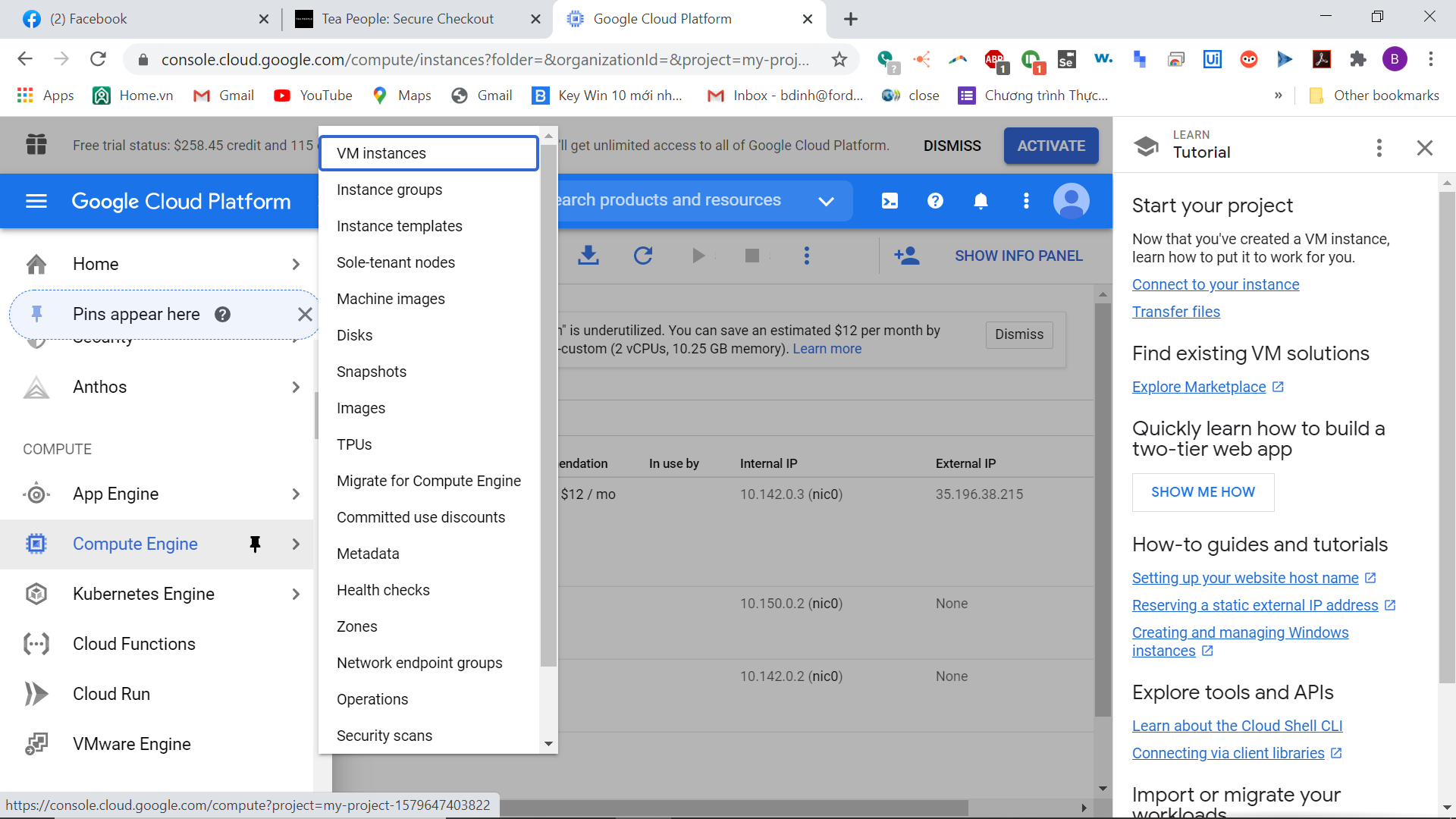


You will see DataIku there, click the three dots next to dataiku vm instance and select **Stop**



## Logging onto DataIku after initial launch

After successfully deploying DataIku on your computer, to launch DataIku again, open the  on Google Cloud and scroll down to **Compute Engine**, and select **VM Instances**

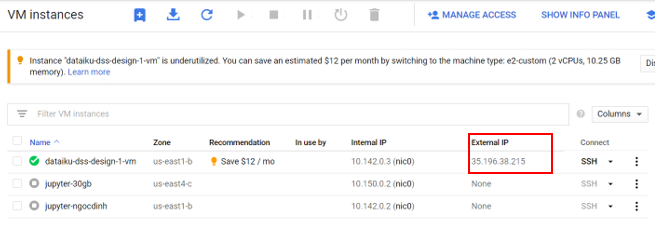


Click on the three dots next to the DataIku instance and select Start/Resume. I already have mine started in this screenshot so it is greyed out.



Then after the DataIku instance has been started successfully, it should have a green checkmark next to the instance.

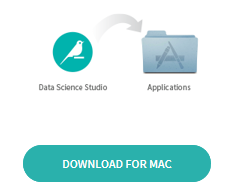
The table also shows the external IP of the instance. Copy the External IP and open it in another tab.



# 3. Install on Mac

Dataiku can be installed directly on Mac OS 10.9 and later. Follow the steps below.

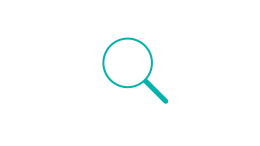
Step 1: Download and install the following dmg: <https://cdn.downloads.dataiku.com/public/dss/8.0.2/Data%20Science%20Studio%208.0.2.dmg>



Double-click to open the DMG file and, when prompted, drop the application into the Applications folder.

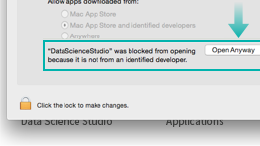
Note: use this same process when upgrading Dataiku DSS to a new version; overwriting the existing application.

Step 2: Start



You can search for the DataScienceStudio application with Spotlight; double-click to start it.

Step 3: Authorize



Depending on your security settings, OS X might prevent DSS from starting.

Open System Preferences, go to “Security & Privacy” and authorize the application.

DSS opens automatically in a browser tab (Chrome and Firefox are supported). If you close the tab and want to reopen DSS, click on the DSS icon in the menu bar and select “Open in browser” or open a new tab and

# 4. GNU/Linux Installation

**Step 1: Download DSS**

wget https://cdn.downloads.dataiku.com/public/dss/8.0.2/dataiku-dss-8.0.2.tar.gz

Or, use this [direct link](https://cdn.downloads.dataiku.com/public/dss/8.0.2/dataiku-dss-8.0.2.tar.gz).

DSS works on Ubuntu, Debian, CentOS, RHEL and Amazon Linux. For version details, please see our Requirements page.

**Step 2: Unpack**

Unpack the downloaded archive where you want to install DSS.

You must keep the directory even after installation is complete.

tar xzf dataiku-dss-8.0.2.tar.gz

**Step 3: Install**

Launch the installation script. You need to choose:

A directory where Dataiku DSS will store configuration and data.

A base TCP port.

dataiku-dss-8.0.2/installer.sh -d DATA\_DIR -p 11000

**Step 4: Start**

DATA\_DIR/bin/dss start

**Step 5: Enter the studio**

Browse to http://<your server address>:11000.

Only Chrome and Firefox are supported.

# 5. Install Using Virtualbox Or VMWare

Dataiku provides a pre-built Linux virtual machine for the free Virtualbox engine. This lets you run DSS for evaluation purpose on Windows or Mac OS

Prerequisites and troubleshooting

There are several prerequisites that you must meet before being able to use the virtual machine. Please read the [Documentation](https://doc.dataiku.com/dss/latest/installation/other/vm.html) carefully. The main prerequisites are summarized below:

* 64 bits CPU with 64 bits OS
* Hardware virtualization enabled in BIOS
* Virtualbox or VMWare player
* 8 GB of memory on your machine

In addition, corporate security suites often prevent proper operation of the virtual machine. See our [Troubleshooting documentation](https://doc.dataiku.com/dss/latest/installation/other/vm.html)

VMWare player support

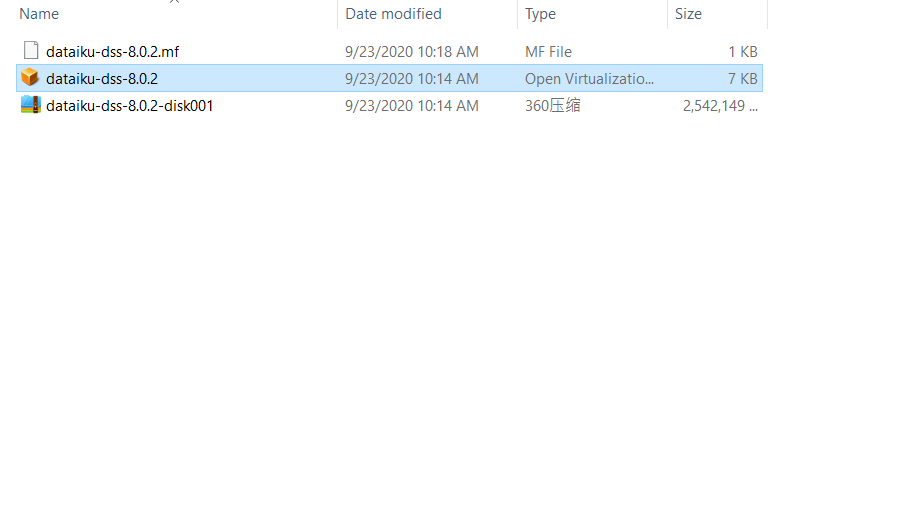
These instructions are also valid for VMWare Player. You will encounter a warning while importing the OVA file. You can ignore it and select “Retry”

**Step 1: Install Virtualbox**

Download and install Virtualbox from [Virtualbox.org](https://www.virtualbox.org/)

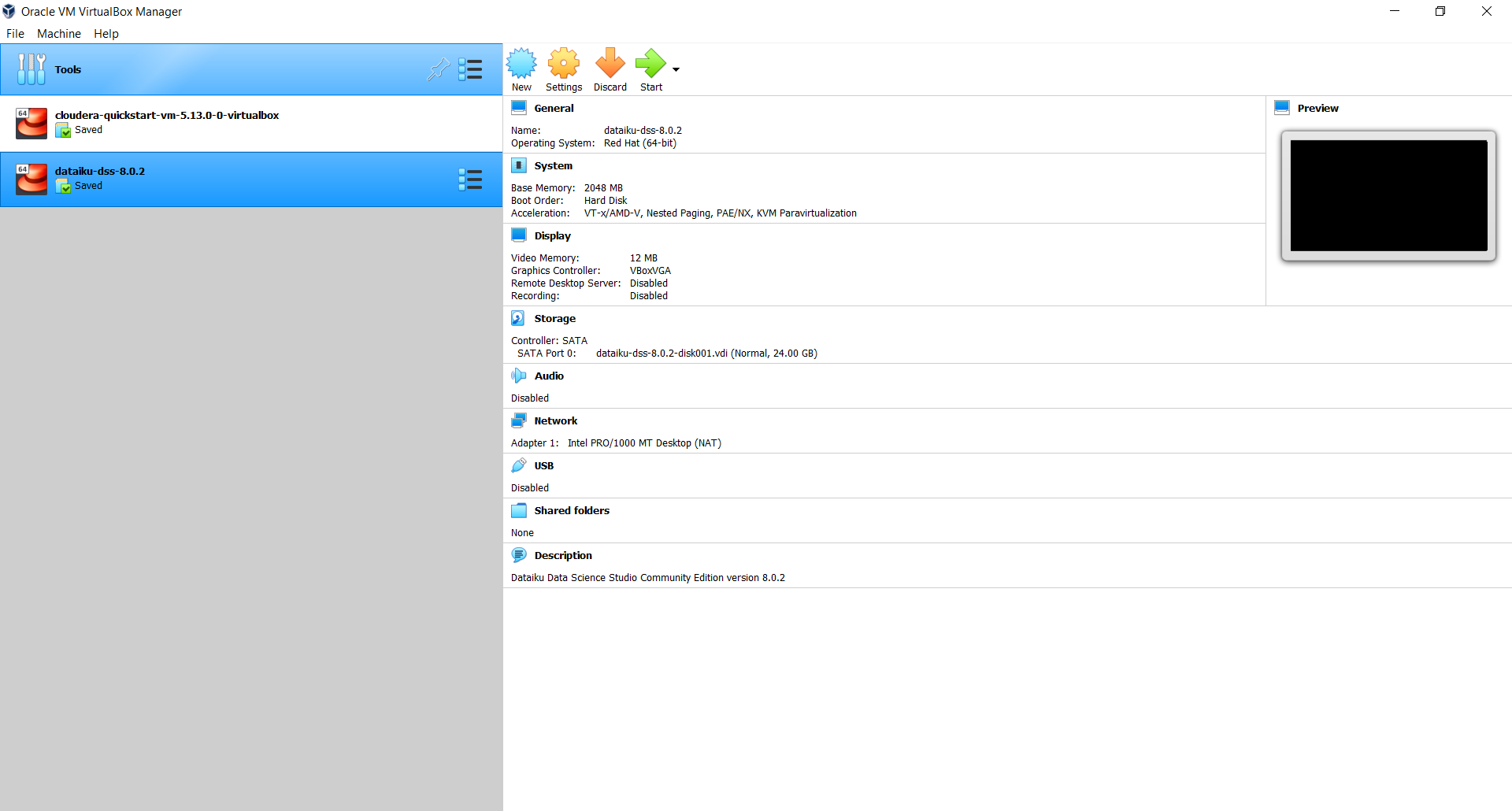
**Step 2: Download the DSS image**

Download the [DSS virtual machine](https://cdn.downloads.dataiku.com/public/dss/8.0.2/dataiku-dss-8.0.2.ova) (2.2 GB).  
Open it in Virtualbox (either by double-clicking or using File > Import Appliance)



**Step 3: Import the Virtual machine**

The Appliance import wizard opens. You should not need to change any setting. Simply click on Import.  
Once the virtual machine is imported, click on Start. The virtual machine boots.

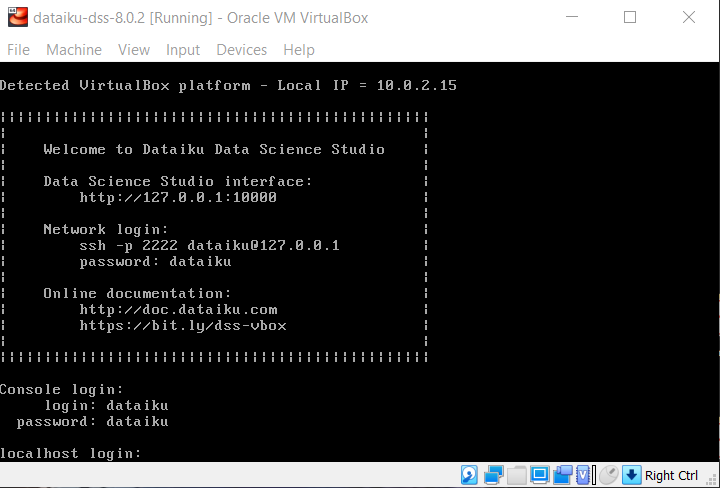


**Step 4: Open DSS**

Once the Virtual Machine has finished booting, it displays a welcome banner with connection details.

You do not need to login into the virtual machine. Open your regular Chrome or Firefox browser (i.e.: not in the virtual machine). In that browser, open the URL that is displayed in the welcome banner, as “Data Science Studio interface”. The DSS interface appears.

This URL to open is often http://127.0.0.1:10000 (but not always, please check the welcome banner).

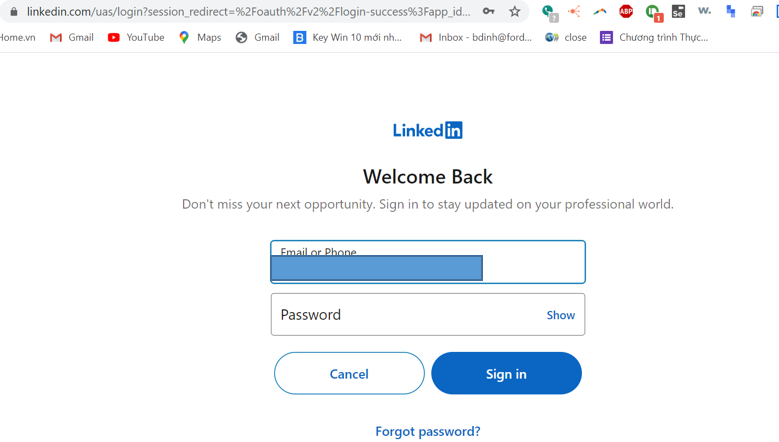


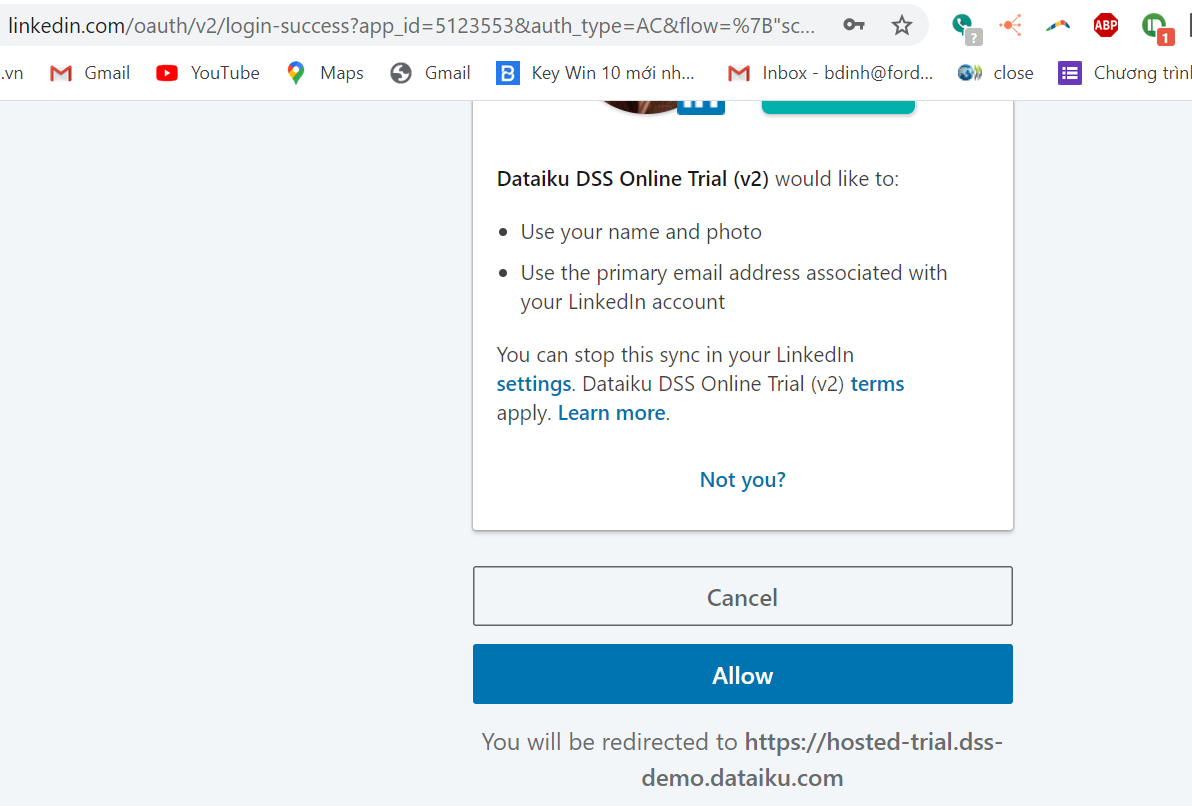
# 6. Hosted Online Trial by DataIku

Note: this should only be used if all other methods failed. Please note that this trial is only for 14 days.

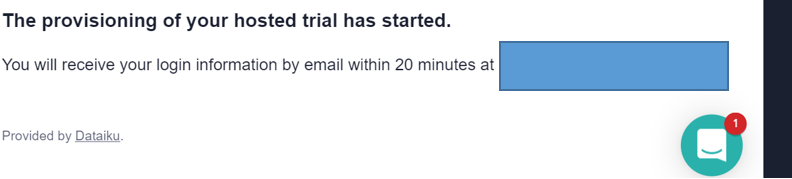
Go to this link: <https://hosted-trial.dss-demo.dataiku.com/linkedinconnect>

Sign in with your LinkedIn credentials.

Click allow



It will take you back to this page.



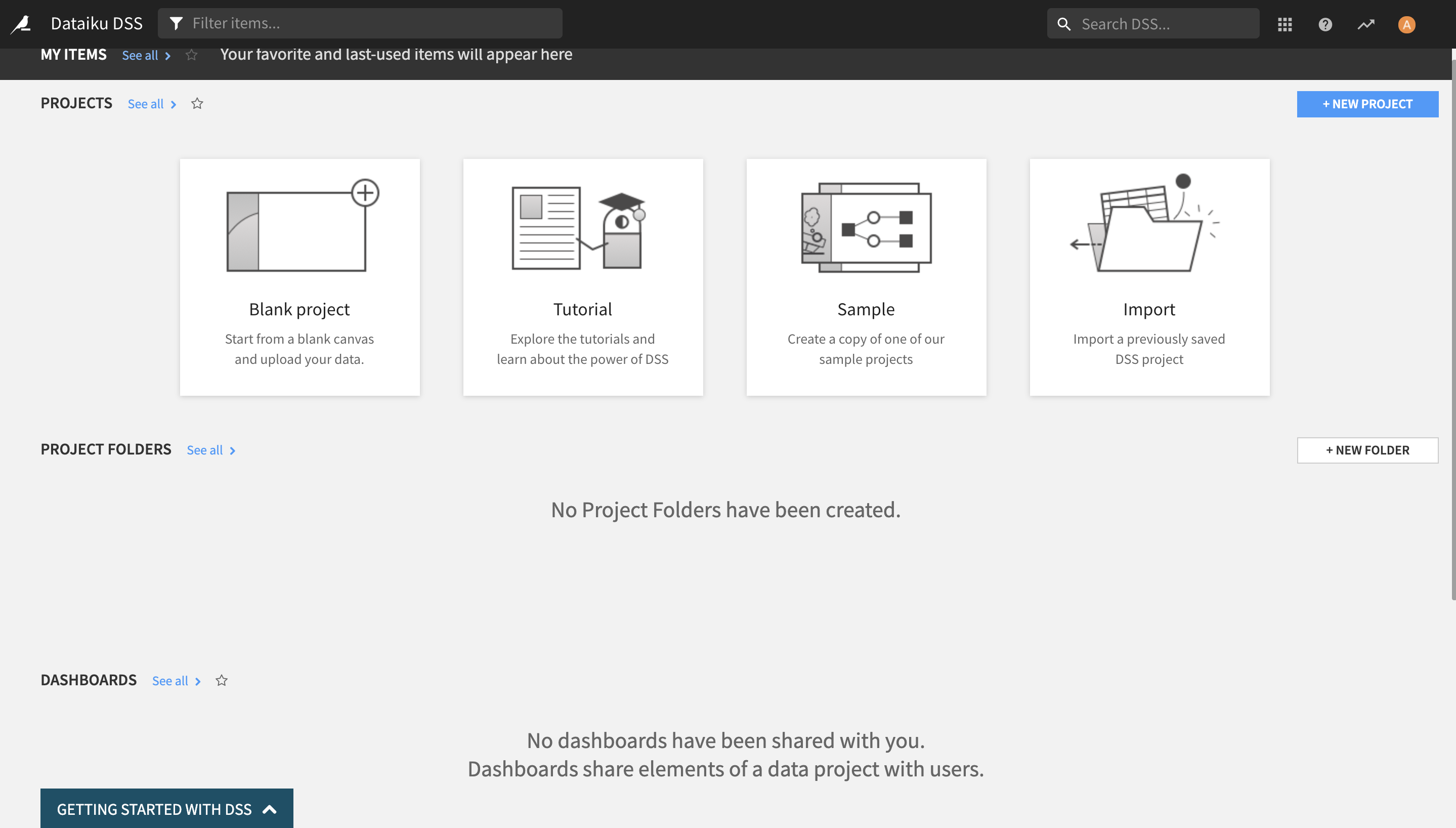
Check your email for credentials and you can get started.

**Part 2 - Create Data Set & Data Exploration**

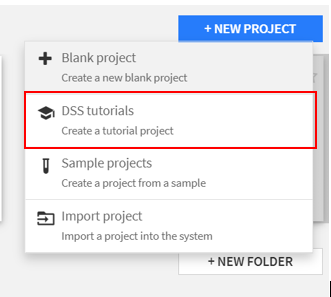
This tutorial mainly introduces the creation of data sets and data exploration. This tutorial uses fuel consumption dataset. This dataset is a fuel consumption dataset, which contains model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada.

# 1.Create a project

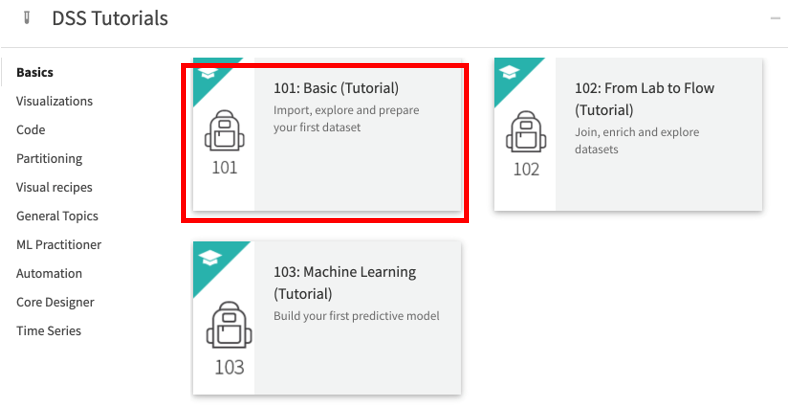
Step 1: First click **Tutorial** to create a project.



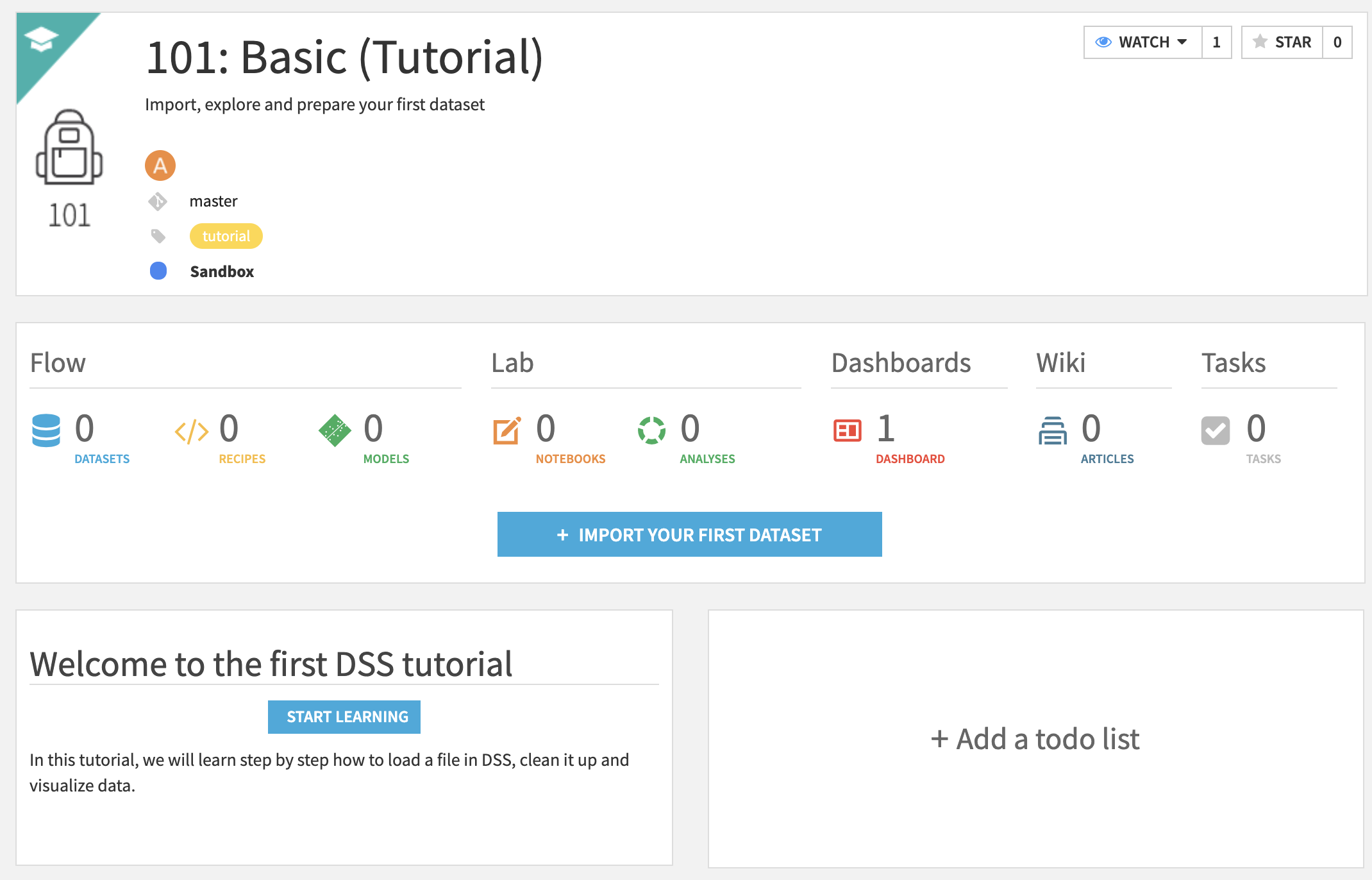
Step 1 (alternative): You can also do this step by click on **New Project**, and select **DSS Tutorials**



Step 2: Create a 101: Basic (Tutorial) project directly from the left column of Dataiku DSS.

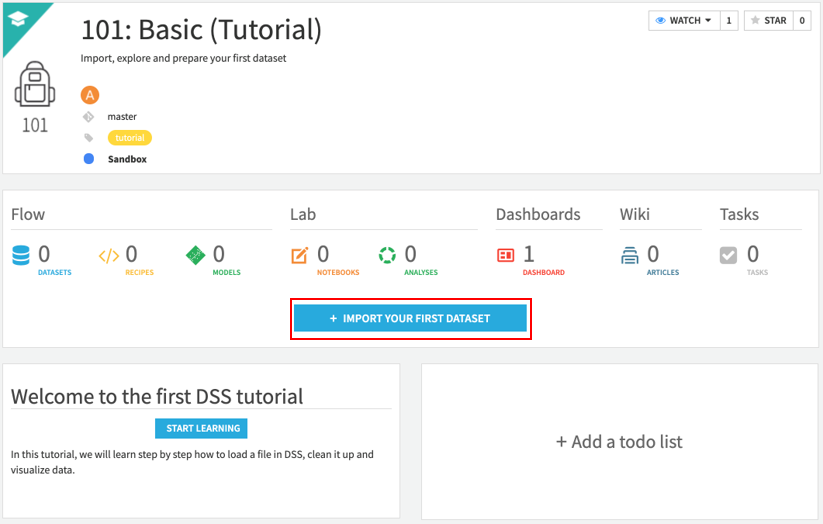


After the creation is complete, the project page is as shown in the figure below.

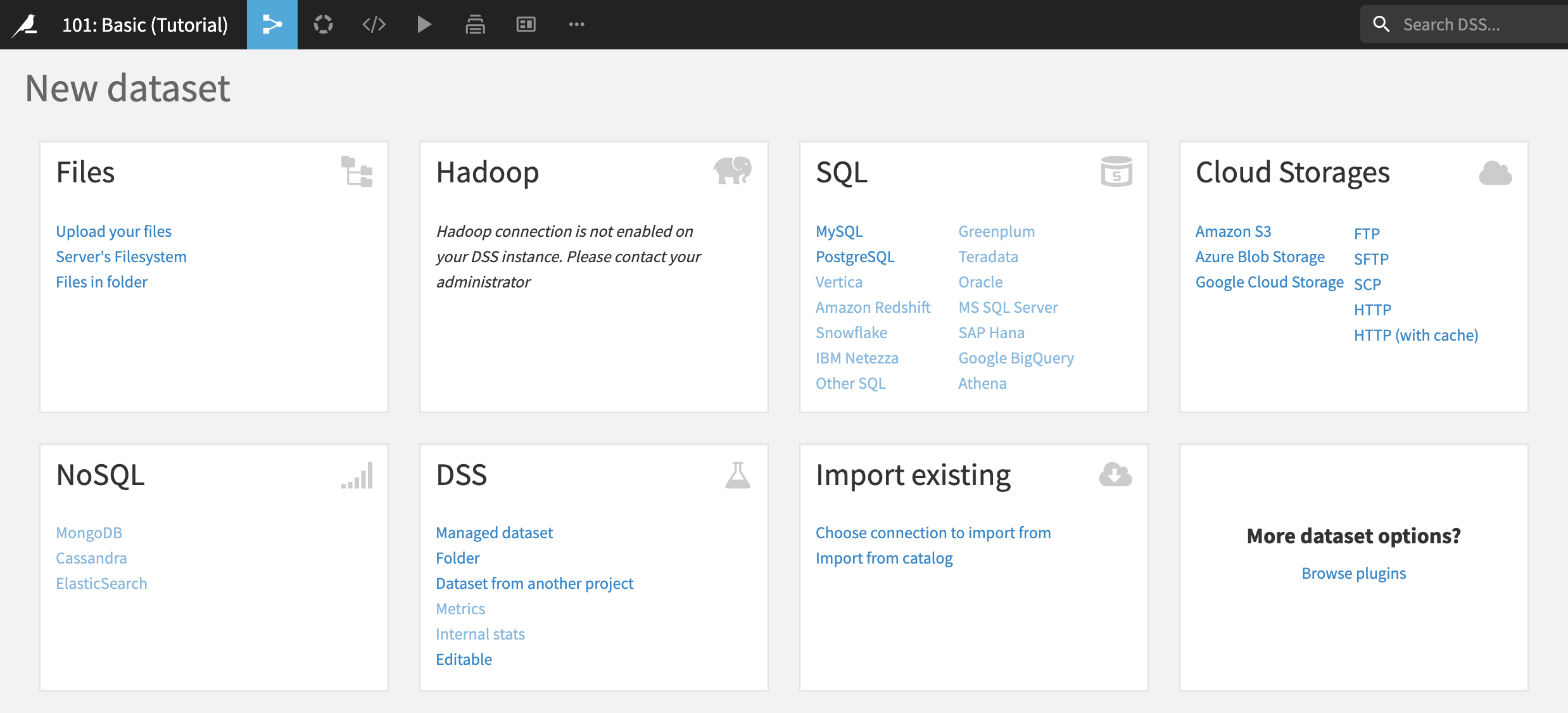


# 2.Create a data set

Step 3: Click **IMPORT YOUR FIRST DATASET** on the project homepage to import a data set.

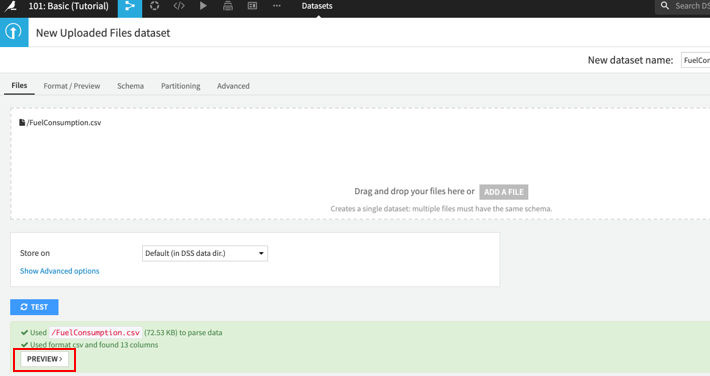
 The figure below shows that Dataiku supports a wide range of data types, and of course there are plug-ins that can be extended. For detailed types and features, please refer to official documents

Step 4: Click **Upload your files** to select the downloaded FuelConsumption.csv file, upload the file to create a data set.



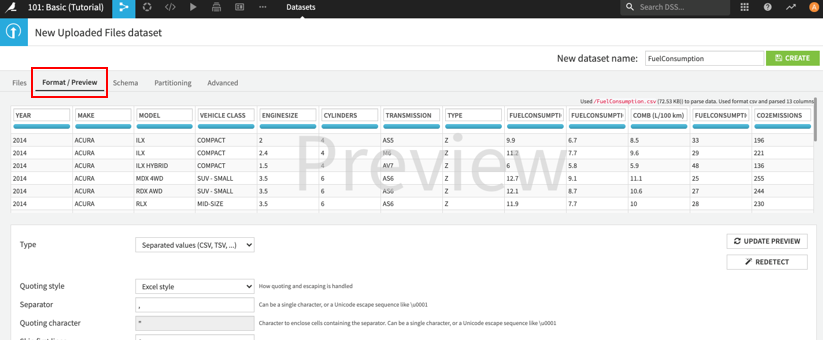
When uploading and creating a data set, you can upload multiple files with the same fields to create a data set together. After the file is uploaded, Dataiku DSS will automatically identify the first non-empty file, identify the file type and fields, and automatically fill in the name of the file as the data set name. In the new version, the data set created by uploading files can be set with filter conditions to filter out which files can be used as data sets; you can also set the storage location, such as HDFS or the server where Dataiku DSS is deployed.

Step 5: Click Preview after successful upload of the dataset.



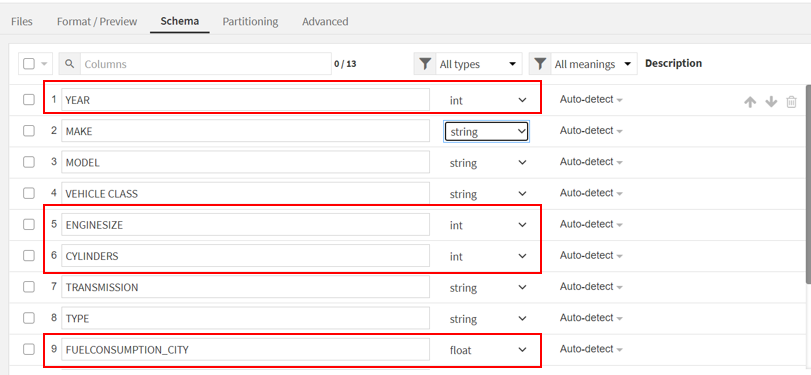
You can configure related parameters on the preview interface.

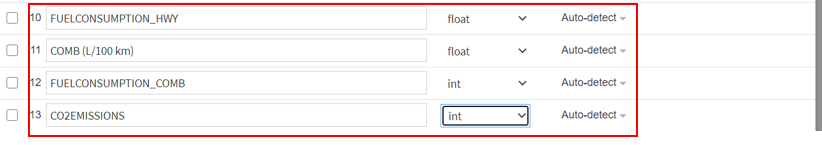
After each configuration option is changed, the preview data will change in real time. Dataiku supports many file types and file-related configurations, including: CSV, Parquet, XML, Excel, MySQL dump, json, etc.

Step 6: Click **Format/Preview** to examine the dataset 

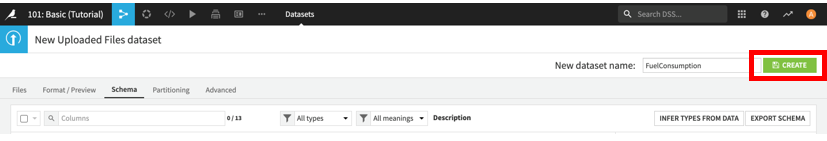
In addition to automatic monitoring success, Dataiku DSS provides the function of modifying fields. Users can add field descriptions, modify the storage type and meaning of fields. meaning is an important concept in Dataiku DSS. It can be seen as a concept strongly related to the data itself and the business itself. The main purpose is twofold: (1) To make it easier for users to understand the meaning of the current field, which is equivalent to helping Document function. (2) Perform verification in accordance with the meaning rules during data exploration and processing to speed up the process of data exploration. In Dataiku DSS, meanings such as text, decimal, integer, boolean, date, object, array, natural language, latitude and longitude, IP address, email address, and User-Agent are provided by default in Dataiku DSS. The following figure shows the meaning provided by the system by default.

Step 7: Click on **Schema** to view the data type that DataIku has already auto-detect for us. If you do not change the data type here, it’s not the end of the world. You can change it later. But I highly recommend changing it here.

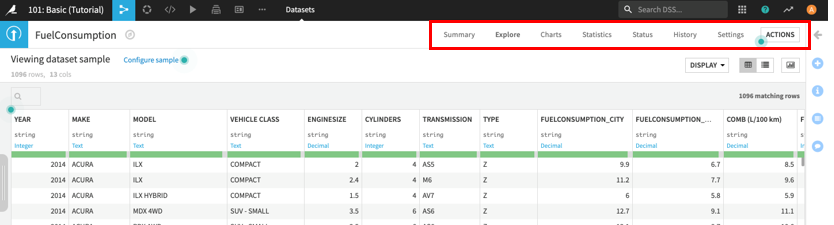




In addition to the built-in meaning of the system, the user can customize the meaning, which can be: declaration (mainly document function), value list, value mapping and regular expression.

Step 8: After all configuration options are confirmed, click **Create** to create the data set. 

Step 9: After the creation is complete, enter the exploration interface of the data set.

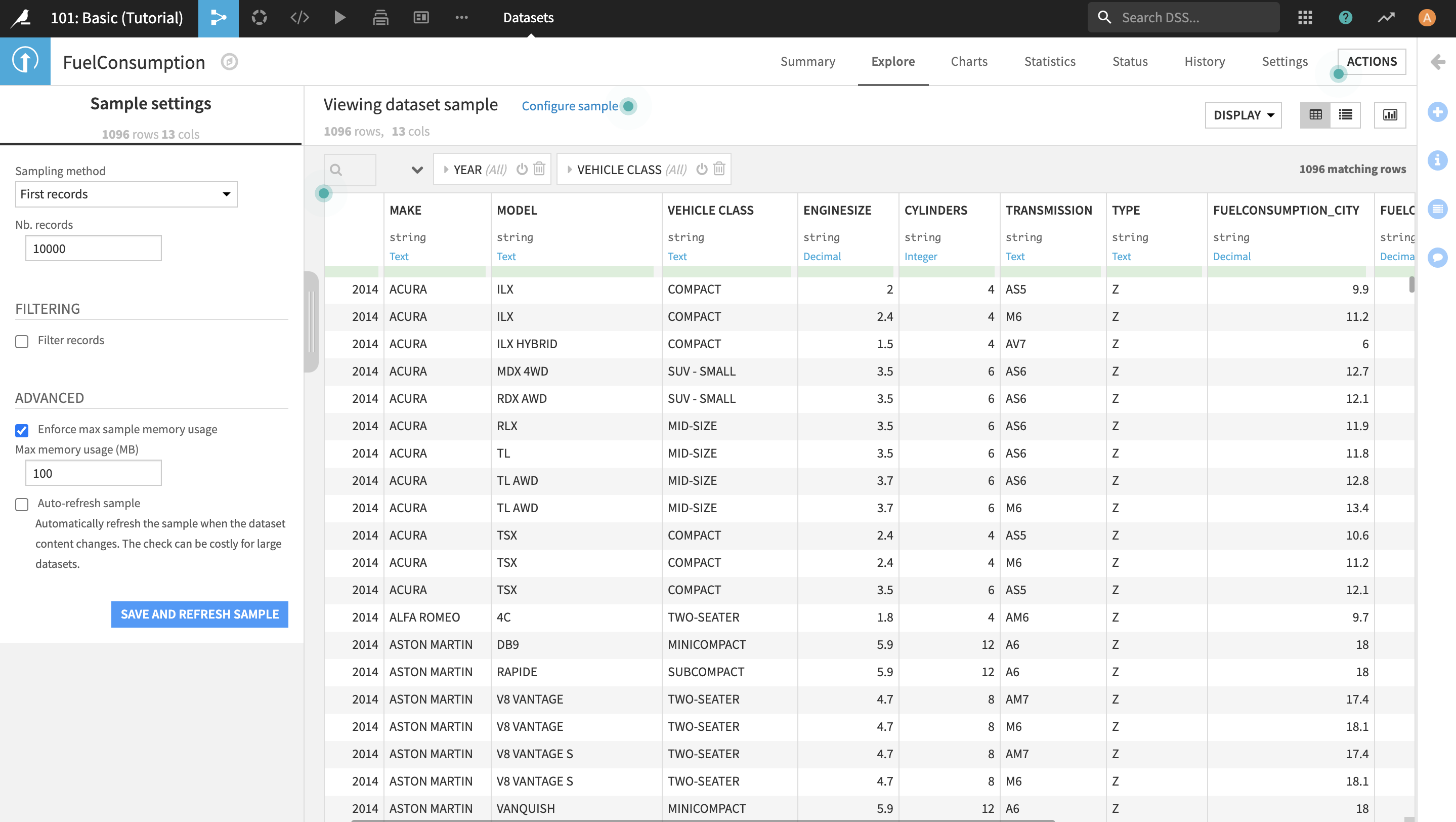


# 3. Data exploration

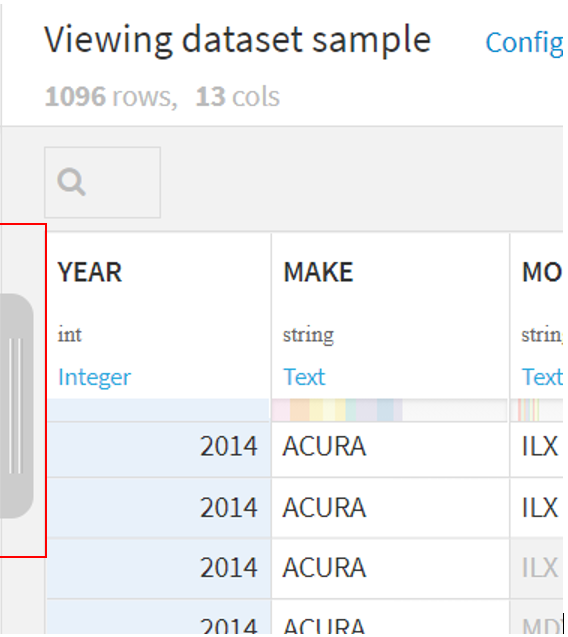
The core functions of the entire data exploration page include: sampling settings, data preview, and data profile.

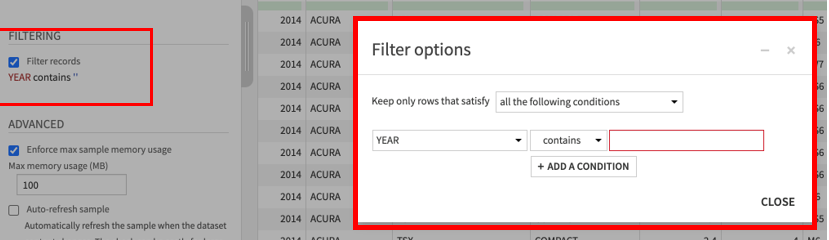
## 1) Sampling settings

The sampling setting is located on the left side of the data exploration interface and provides a wealth of data sampling methods, including fixed number of data sets (random, first N, last N) sampling, fixed ratio sampling, stratified sampling, cluster sampling, etc. Of course, for the partitioned data set, there are options to choose which partitions to sample in.



Note: if the sample settings pane is hidden for you, you can open it by clicking the following:

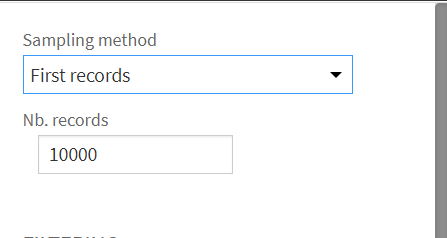


Dataiku considers the problem of data filtering when sampling, so there is a filter option in the sampling setting interface, which can add rich filtering conditions, allowing users to filter faster and extract appropriate data. The following figure shows the configuration options for data filtering: 

In addition to conditional filtering, the sampling setting also provides an option for the size of memory available for sampling data. The maximum memory used by the current data set sampling data can be limited, and the default is 100M. After setting the above options, you can click SAVE AND REFRESH SAMPLE, and Dataiku DSS will resample together. In addition, the system provides the option of real-time resampling, but this option is turned off by default. It should be considered for performance.

To sample data based on first record, filtering on year 2014, we would do the following steps:

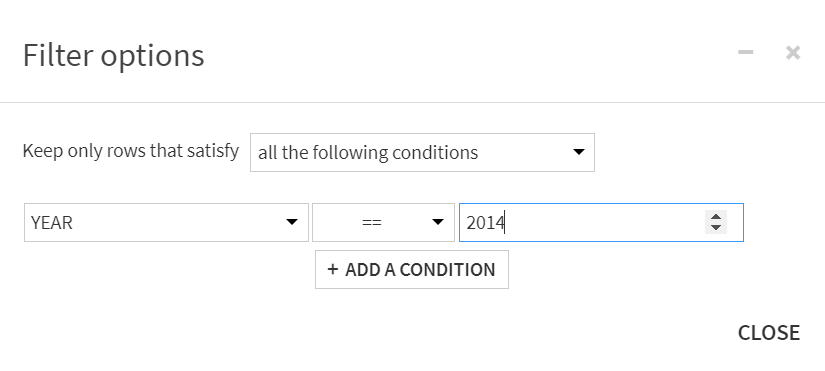
Step 10: Keep **Sampling method** as **First Records** and **Nb. Records** as **10000**



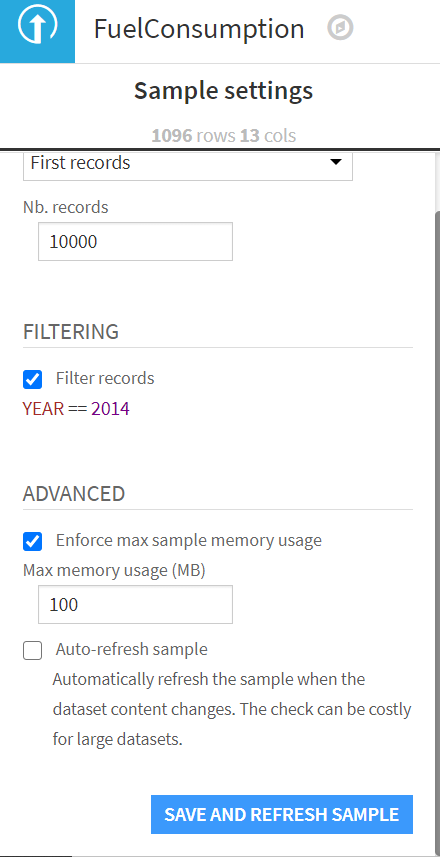
Step 11: Tick **filter records**



Step 12: Enter the following information in the pop-up screen and click **Close**



Step 13: Now that information is filled, the sample settings should look something like this. To save, click Save and Refresh Sample. **But this time we are not going to save.**



## 2) Data preview

The data preview occupies the right part of the entire page. The way of dynamically loading data provides users with the possibility to preview more data faster. The user can always pull down the scroll bar quickly, the interface will briefly display the wait state, but the current row and column data will be loaded soon.

In order to facilitate the exploration of the data, the data preview provides two color viewing modes: meaning verification mode and value distribution mode.

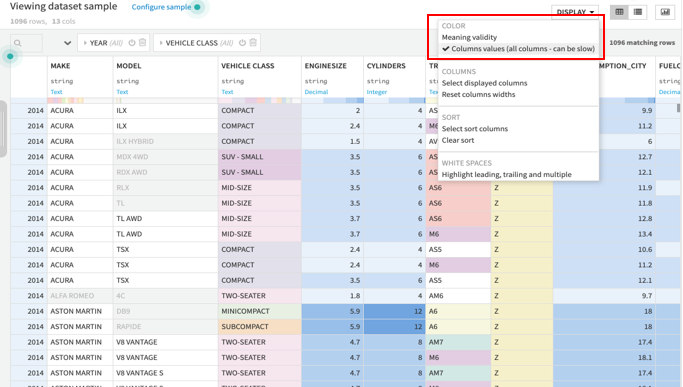
### Meaning verification mode

In this mode, the data preview page will display: the field name of the data set, the storage type of the field, the meaning of the field, the status bar and the data of the field. Dataiku will verify the cells in the current column that meet the meaning rules according to the meaning of the field. If the content of the cell does not meet the meaning of the current column, it will be displayed in red shading. The status bar of the field contains three parts of information: the proportion of non-empty and validated values ​​in the field (green), the proportion of empty values ​​in the field (gray), and the proportion of values ​​that do not meet the meaning rule in automatic (red). This colored status bar allows users to quickly see the data status.

### Numerical distribution model

In this mode, Dataiku DSS will perform binning statistics on certain fields (the specific rules for selecting which columns and how to bin are not very clear, but ID and date will not be included in binning), and statistics the following distributions of the current sampling data, and Each bin in the column is assigned an asymptotic color. Similarly, the shading color of the cells in the data table will also be represented by this asymptotic color. As shown below:

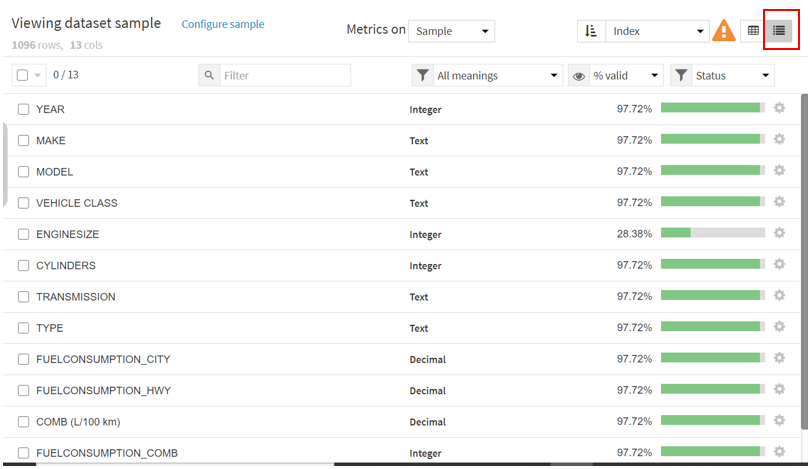
Step 14: Click on **Actions**, select **Column values (all columns – can be shown)** under **Meaning Validity**



Parallel to the color mode, Dataiku DSS provides the option of selecting display columns. In the case of particularly many fields, users can choose to display certain columns. In addition, Dataiku provides the ability to sort and view data. Users can choose which columns to use and which way to preview the data.

## 3) Explore Profile

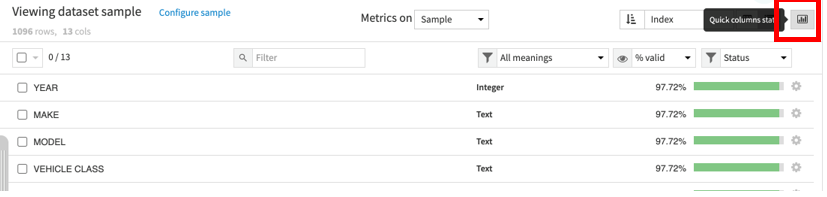
In addition to the provincial preview of data, Dataiku DSS provides a convenient data profile capability.

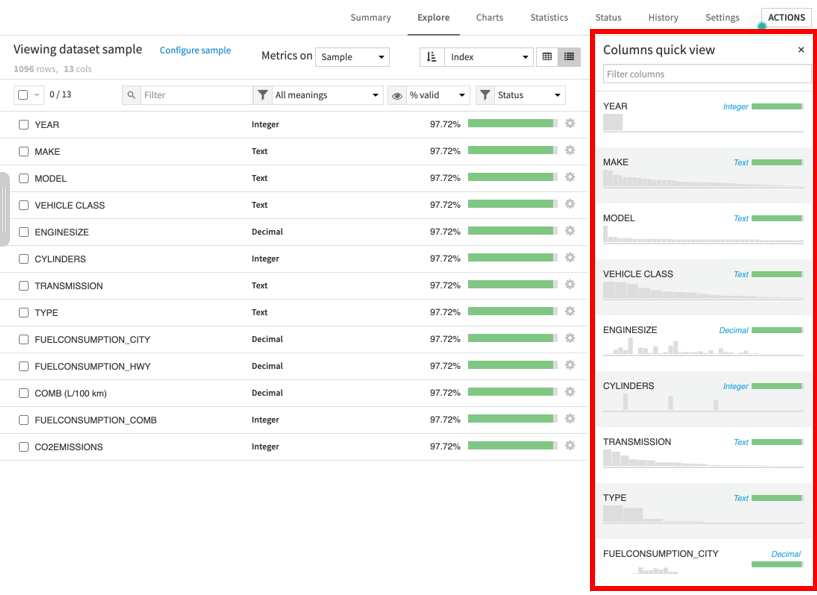
Step 15: Click **Columns View** in the upper right corner to enter the profile page of the data. 

On this page, you can see the meaning type of each field, Profile indicators and the status of the field. Profile indicators include two types. One type can be used for both string and numeric types, such as Cardinality, the number of rows that meet the meaning rules, the number of rows that do not meet the meaning rules, null values, non-null values, and modes. The other is only valid for numeric fields, such as maximum, minimum, average, standard deviation, sum, and median. The status of the field is still indicated by the status bar.

The profile of the data is the profile based on the sampling data by default. Of course, the profile can be made on the entire data set. The Metric on at the top can switch the sampling data and the full data of the data set.

In addition to this Profile information, Dataiku provides a quick profile capability on the right.

Step 16: Click Quick columns stats to quickly see the histogram of each column. Hover the mouse over the histogram, you can column the number and proportion of each value. 



This article mainly introduces the most basic data exploration capabilities of Dataiku DSS. In fact, there are also Chart and data set status alongside data exploration. It can also be regarded as another way of data exploration.