IBM WDP Quick Start Guide + Data Science Experience Hands-on

The purpose of this document is to have a quick start guide on the platform.

This guide does not cover all the features at all, it's just an introduction.

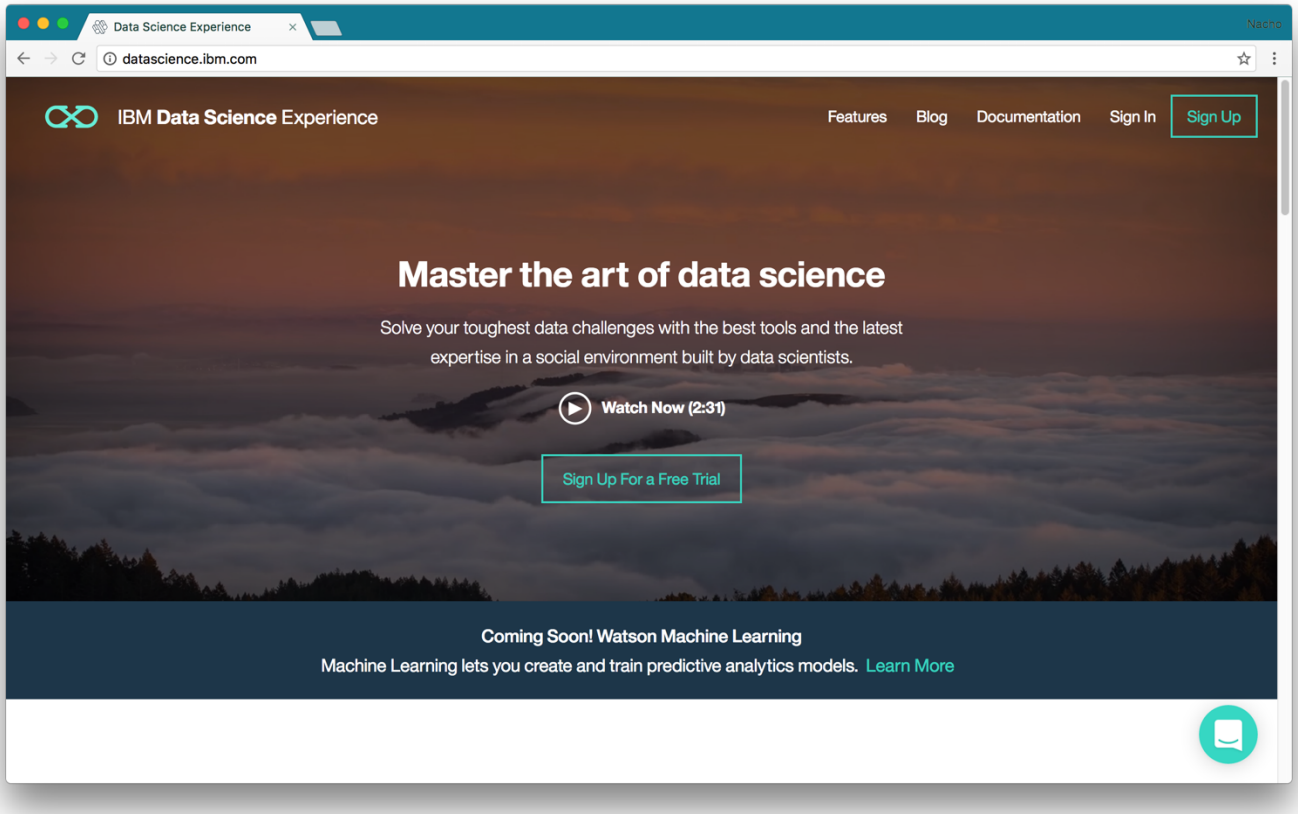
Table of contents

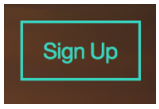
1. Registration on Data Science Experience (DSX)

Enter on a web browser. (It's recommended to use Google Chrome).

Type the following URL: datascience.ibm.com (or just click on the previous link).

You should be a welcome page like this one:

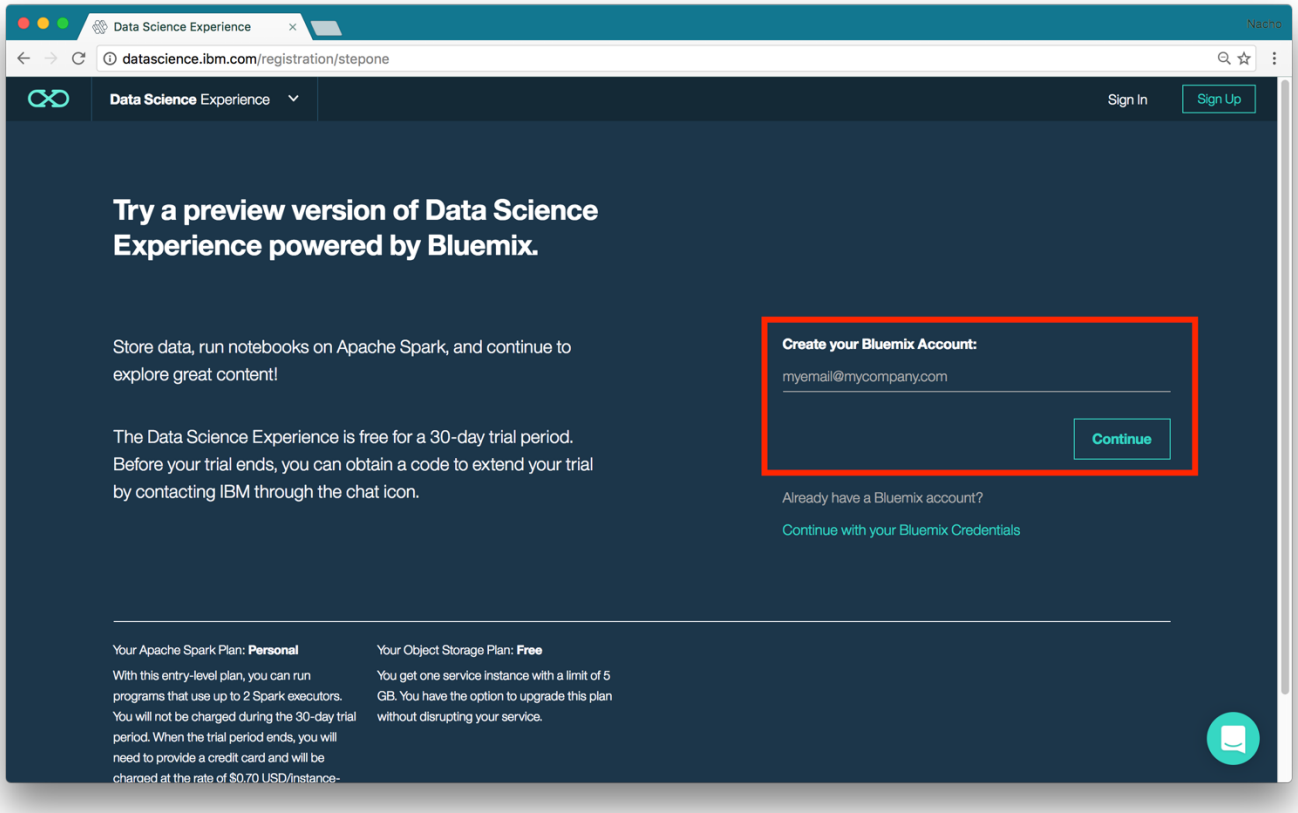


Click on the **Sign Up** button: 

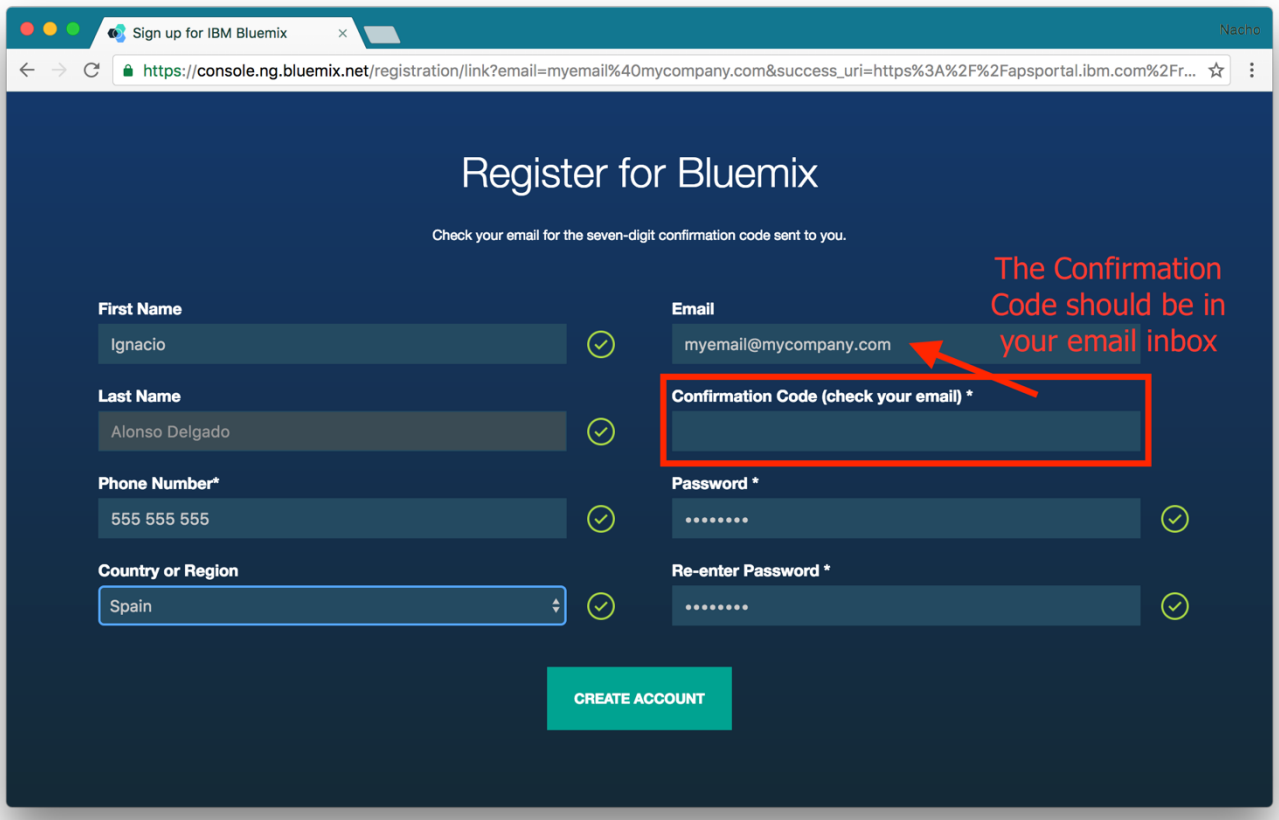
Most of the services that Watson Data Platform uses, are in Bluemix. So now, there are two options, if you don't have a Bluemix account, go to next point (1.1), if you already have a Bluemix account go to 1.2,

1.1. First time using Bluemix

If this is your first time using Bluemix and you don't have an account yet, create a new account following the form. Please, make sure that you enter your email correctly, because the system will send a **Confirmation number** code to this email.



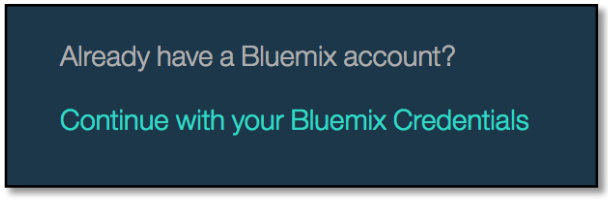
On the next screen, enter your info. And check your email (entered on the previous screen) to copy the **Confirmation Code**:



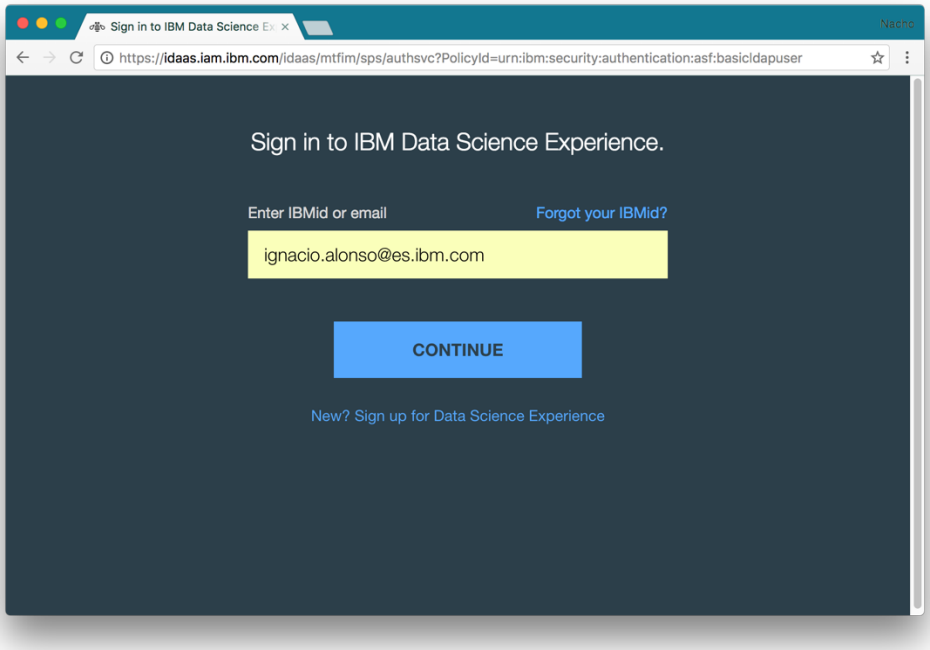
Then, create a new Organization and Space in Bluemix.

1.2. I already have a Bluemix Account

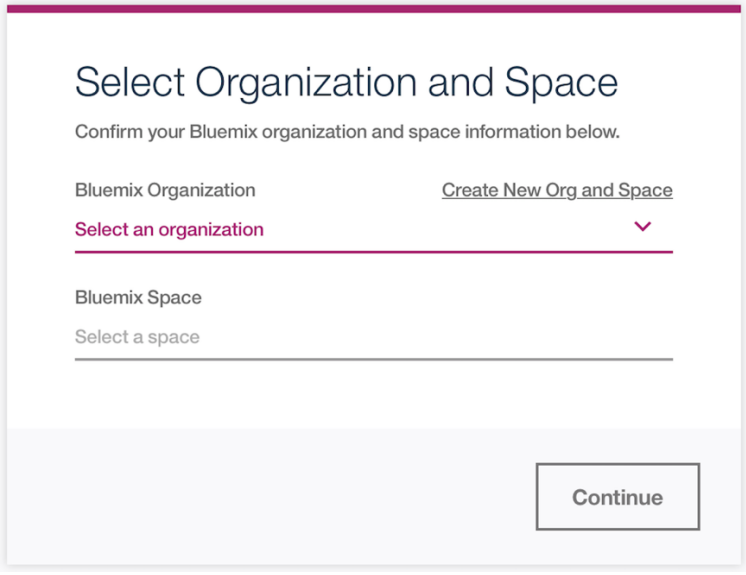
If you already have a Bluemix account, just click on "Continue with your Bluemix Credentials":



Then, enter your Bluemix credentials (email & password):

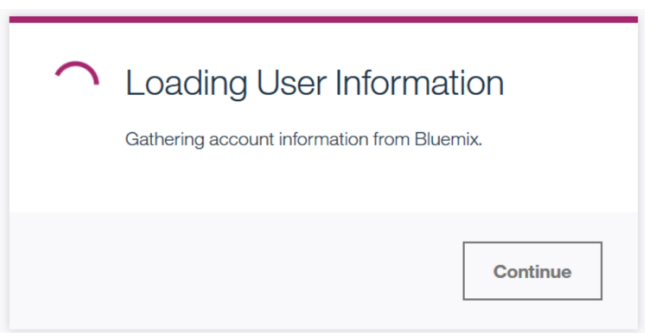


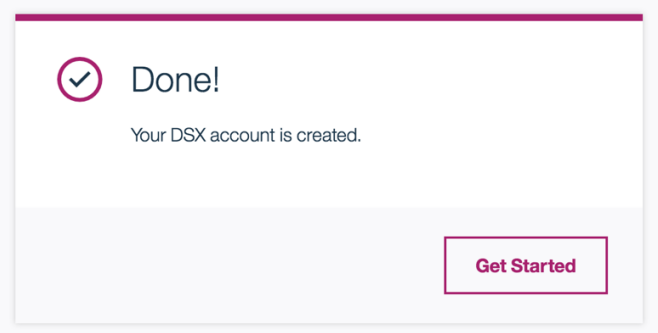
If the next pop-up window appears, select your Bluemix Organization and Space. Click on Continue:



2. Discovering DSX – Data Hub

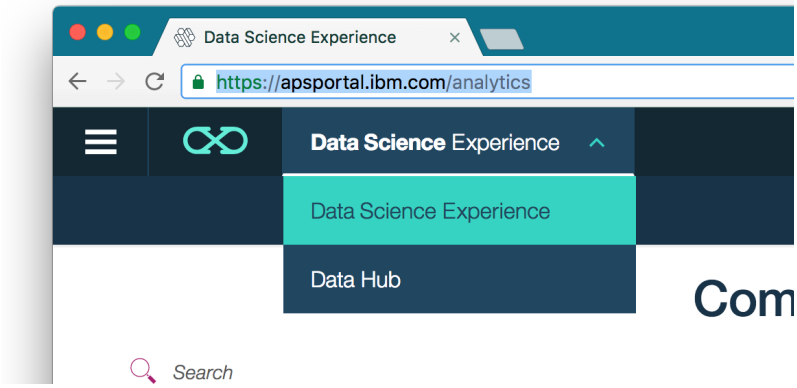
Once you have successfully login, you should see these two pop-up messages:





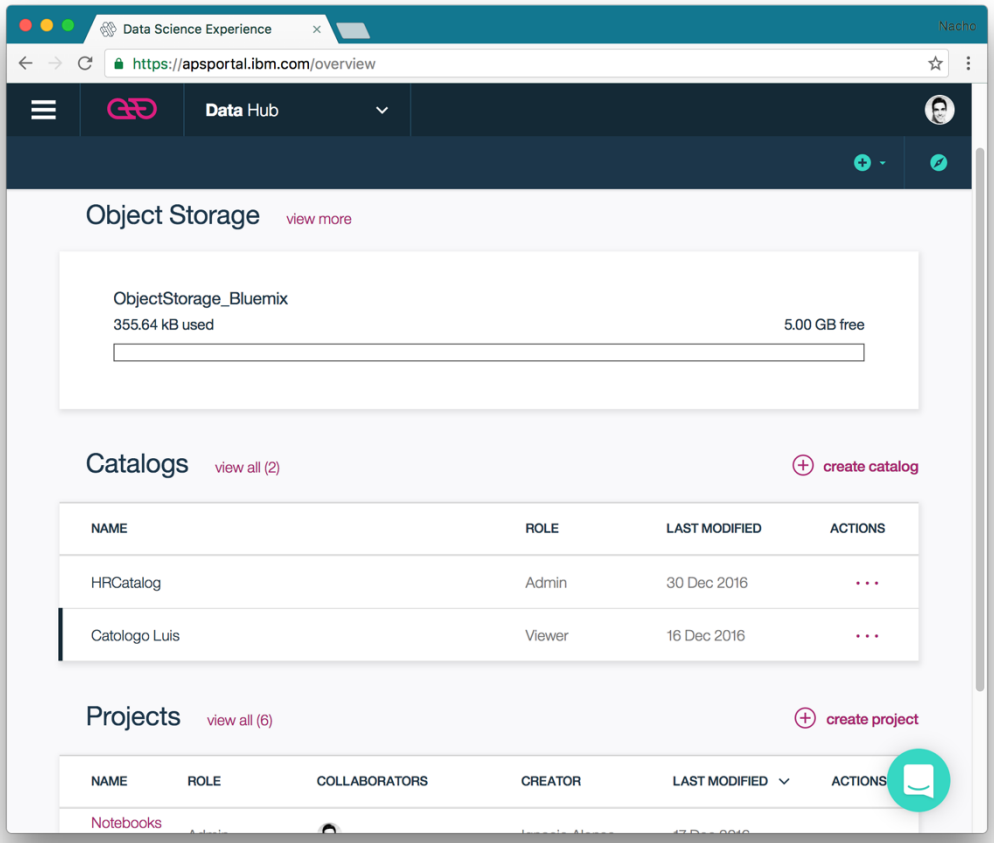
Click in "Get Started" button.

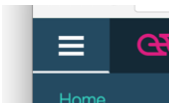
On the top menu, you can switch between different environments. Click in "**Data Hub**" to explore it.



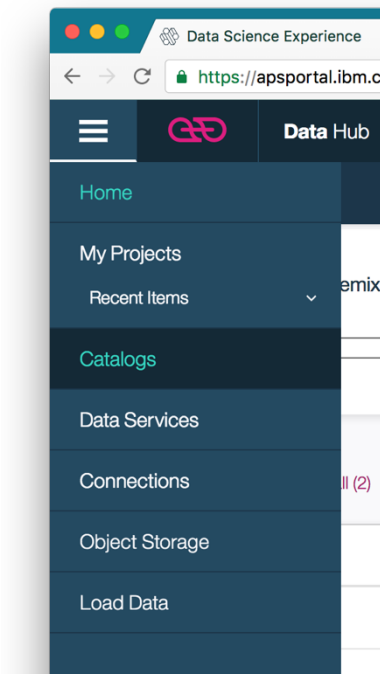
This will show you the welcome page of Data Hub (next screenshot). Here you have access to:

- **Object Storage**: You have 5GB of Object Storage space for free in the free account.
- **Catalogs**: A catalog is a cloud-based data repository for structured and unstructured data from a variety of sources. You copy and categorize data assets into the catalog and add collaborators who can analyze the data. You create projects to organize data, analytical tools, and collaborators. A catalog has two types of storage areas for data assets:
    - *Shared zone*: Contains governed, read-only data assets. All catalog members can see all data assets.
    - *Project sandboxes*: Contains updatable data assets that belong to a project. Each project has a separate sandbox. Only project members can see and edit project assets.
- **Projects:** If you associate a project with a catalog, you can add catalog assets to the project. Then, project members can include catalog data assets in their notebooks. You can edit data sets and save them in the project sandbox.



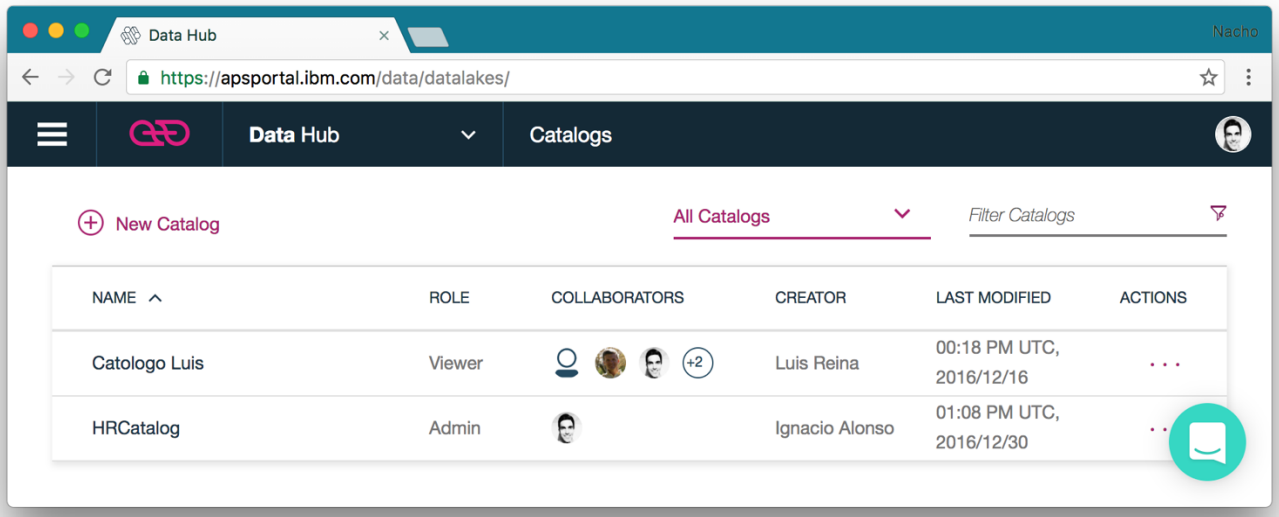
In the upper left corner, you will find a contextual menu  that will change depending on the environment in which you are (nowadays, there are two: "Data Hub" or "Data Science Experience"). Click on it.

And then, click on **Catalogs**.



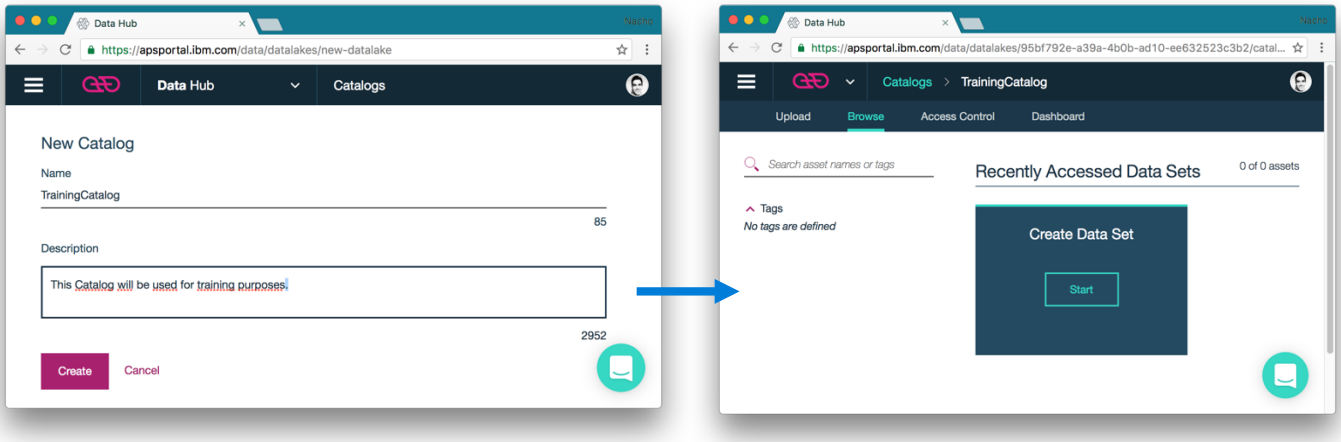
3. Creating a new Data Hub Catalog

If you followed the previous point (2 Discovering), you should be on the **Catalogs** screen. From here, you can see what catalogs you have, and the catalogs that others are sharing with you. You can also see the *Collaborators*, the *Creator*, and the *Last Modified* date.

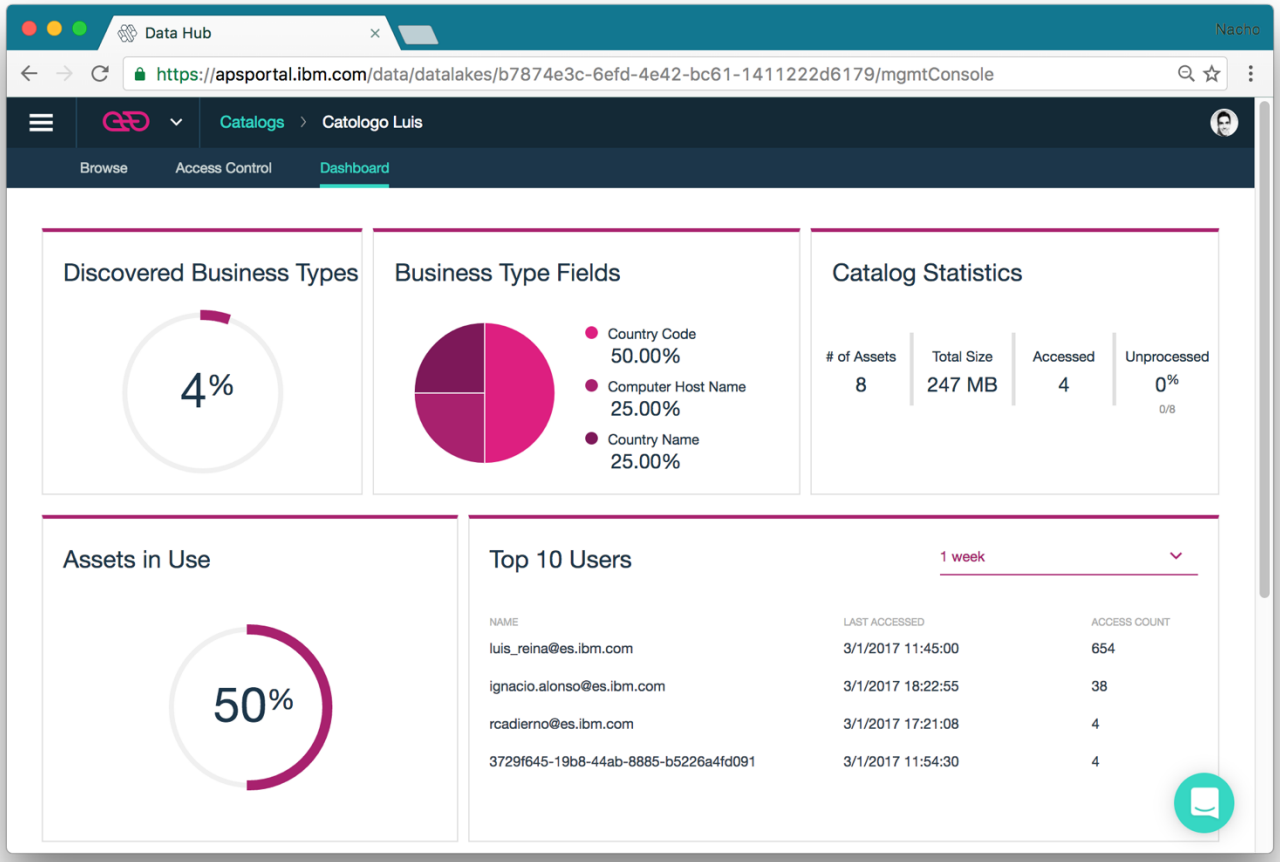


Click on "+ New Catalog" button 

Write a new name (for example TrainingCatalog) and a short description. Click on "Create"



Click on the "Dashboard" tab. As you have not yet uploaded or connected any data, you will see this screen empty. However, you can see an example dashboard below. In it you can see business information about the type of data of this catalog, the top 10 users, the assets being used and other statistics.


Data Hub
Nacho
https://apsportal.ibm.com/data/datalakes/b7874e3c-6efd-4e42-bc61-1411222d6179/mgmtConsole
Catalogs
Catologo Luis
Browse
Access Control
Dashboard
Discovered Business Types
4%
Business Type Fields
Country Code
50.00%
Computer Host Name
25.00%
Country Name
25.00%
Catalog Statistics
# of Assets
8
Total Size
247 MB
Accessed
4
Unprocessed
0%
0/8
Assets in Use
50%
Top 10 Users
1 week
NAME
LAST ACCESSED
ACCESS COUNT
luis_reina@es.ibm.com
3/1/2017 11:45:00
654
ignacio.alonso@es.ibm.com
3/1/2017 18:22:55
38
rcadierno@es.ibm.com
3/1/2017 17:21:08
4
3729f645-19b8-44ab-8885-b5226a4fd091
3/1/2017 11:54:30
4

3.1. Uploading data to the Catalog

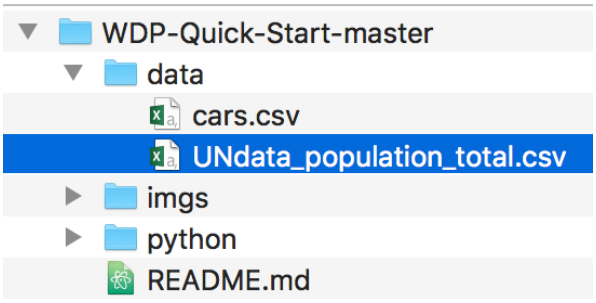
We are going to upload data in our catalog. To do this, the first thing is to have a sample data set.

We will use this **GitHub** repository: https://github.com/nachoad/WDP-Quick-Start

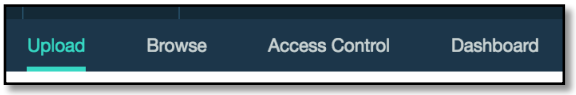
Click on the URL to download the .zip with the content:

https://github.com/nachoad/WDP-Quick-Start/archive/master.zip

Decompress the zip on your laptop. You should see a structure like this:

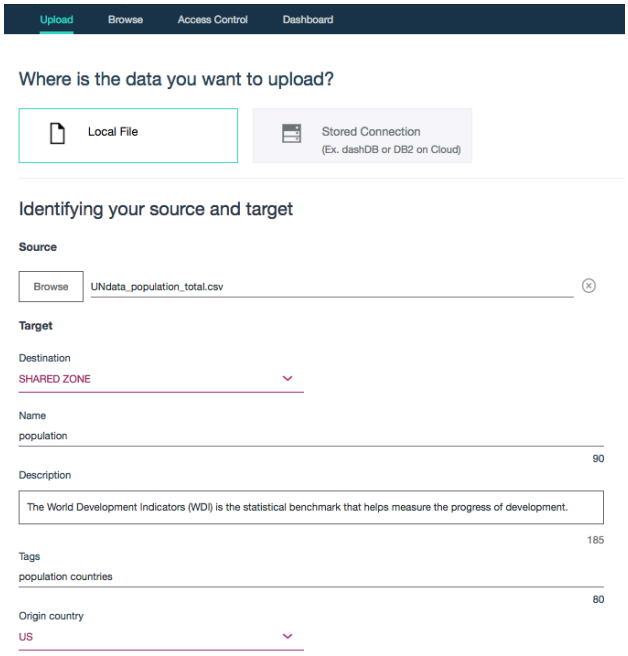


Ok, now we have sample data: `UNdata_population_total.csv`. Go to your Catalog "TrainingCatalog" > Upload tab.

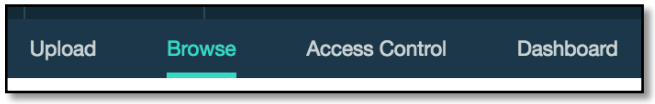


Click in Local File, and click in Browse. Find the `UNdata_population_total.csv` file you just downloaded and select it as Source.

Select Destination: SHARED ZONE. Name: **population**. Enter a short description. Enter some tags, and select the Origin country: US. Click on Upload button. 

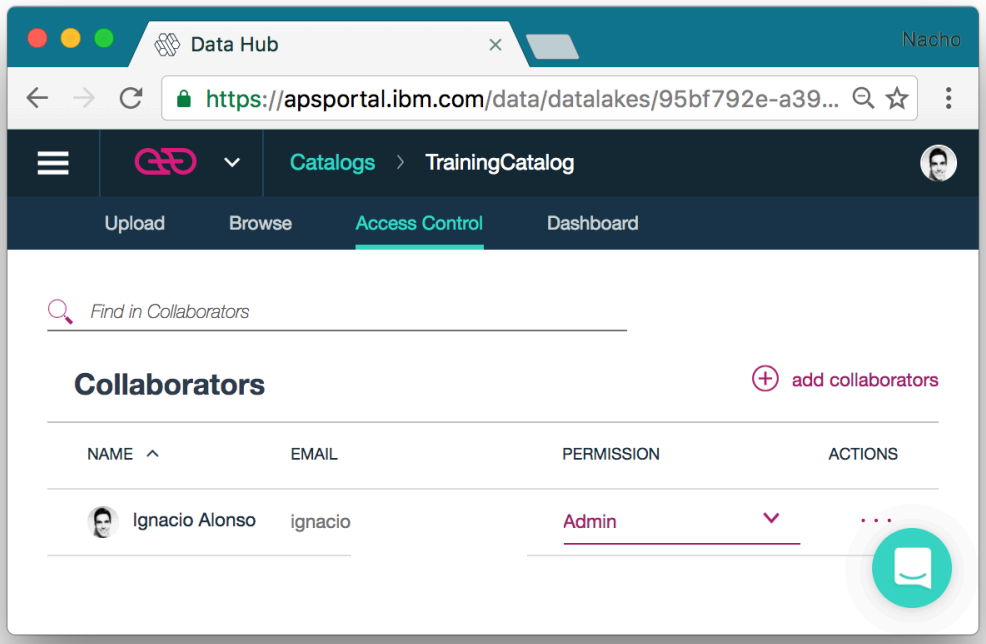


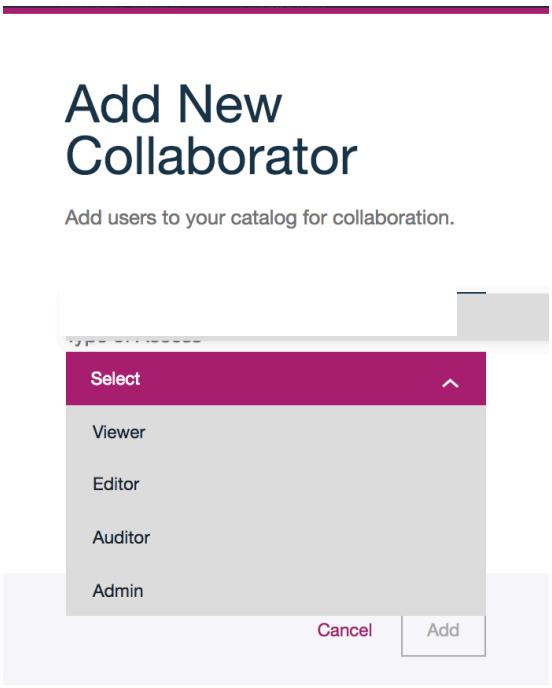
Go to the Browse tab to see the data uploaded.



3.2. Sharing the Catalog with Colleagues

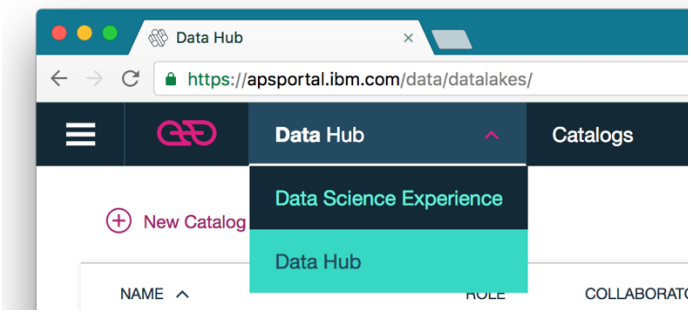
Go to Catalogs > "Training Catalog", then select the "Access Control" tab and click on the "add collaborators" button.





4. Discovering Data Science Experience (DSX)

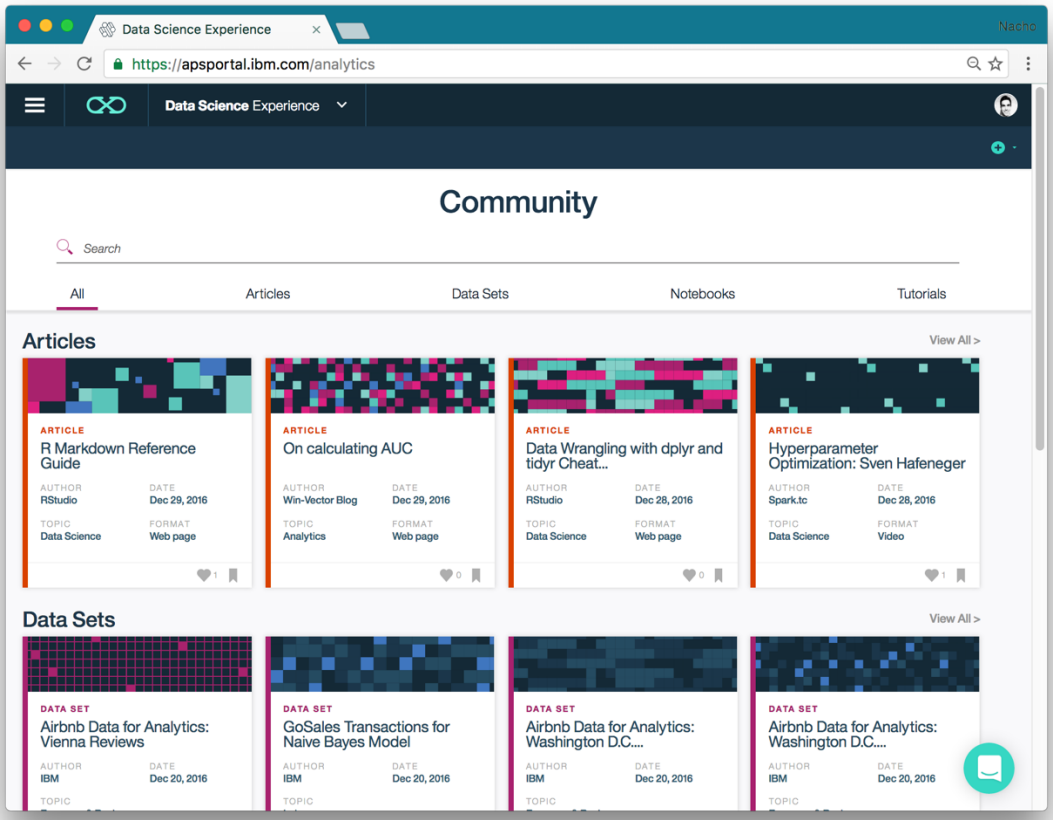
On the environment switch, change to Data Science Experience.



The following screenshot shows the Data Science Experience (DSX) home page. From here you can access to:

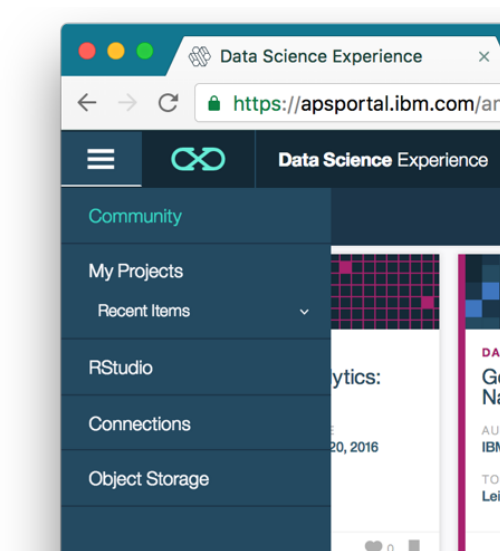
- **Articles**: Documents (blogs, news, articles…) written by IBM people or external, related to the world of analytics, data science, predictive models, etc.
- **Data Sets**: Data sets in various formats (CSV, JSON, DBs, etc.) obtained from free data, from other studies, etc. You can use these data sets for free to start with a new course, study, or tutorial.
- **Notebooks**: A Jupyter Notebook is a web-based environment for interactive computing. You can run small pieces of code that process your data, and you can immediately view the results of your computation. You can create a notebook in R, Python, or Scala. Start from scratch, import an existing notebook, or use one of the samples from the community. Notebooks include all of the building blocks you need to work with data:
    - The data
    - The code computations that process the data
    - Visualizations of the results
    - Text and rich media to enhance understanding

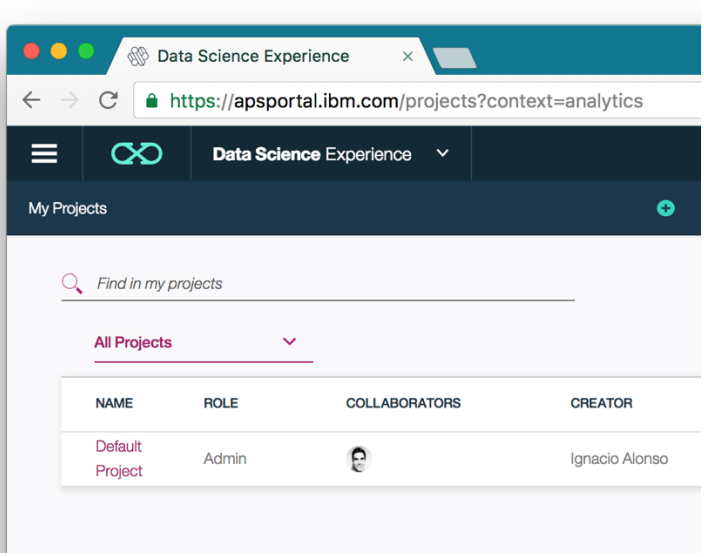
  Here you can see notebooks (ready-to-use) related to different topics like Society, Science, Transportation, Economy, Business, Health, etc.
- **Tutorials**: A good way to start learning different topics like: Spark, Python, R, how to analyze Data Sets, …

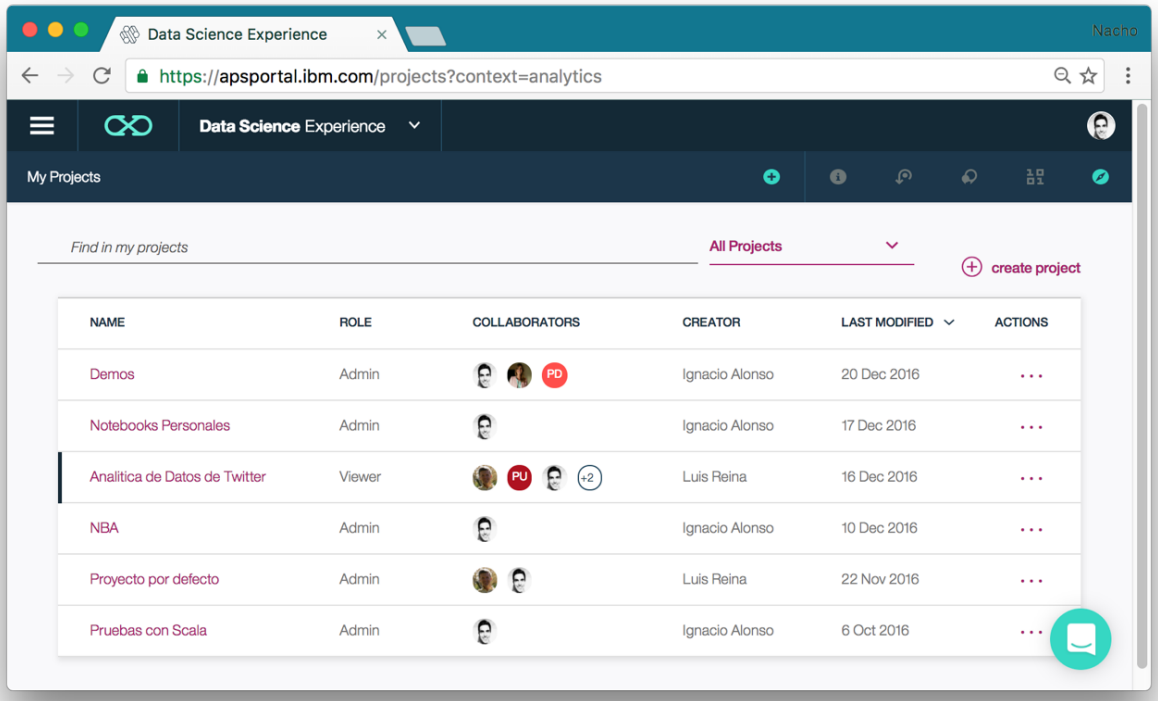


4.1. Creating a new Project

Click on the upper left menu, then click on "My Projects". When creating the new DSX account, a "Default Project" is created. But, let's create a new one.



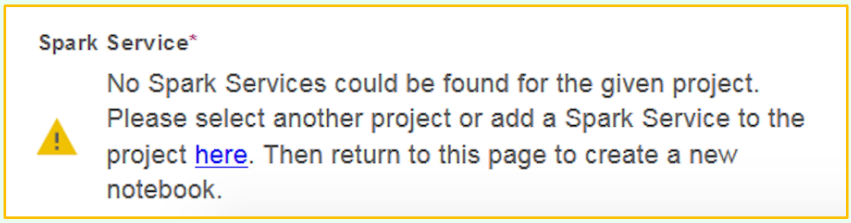




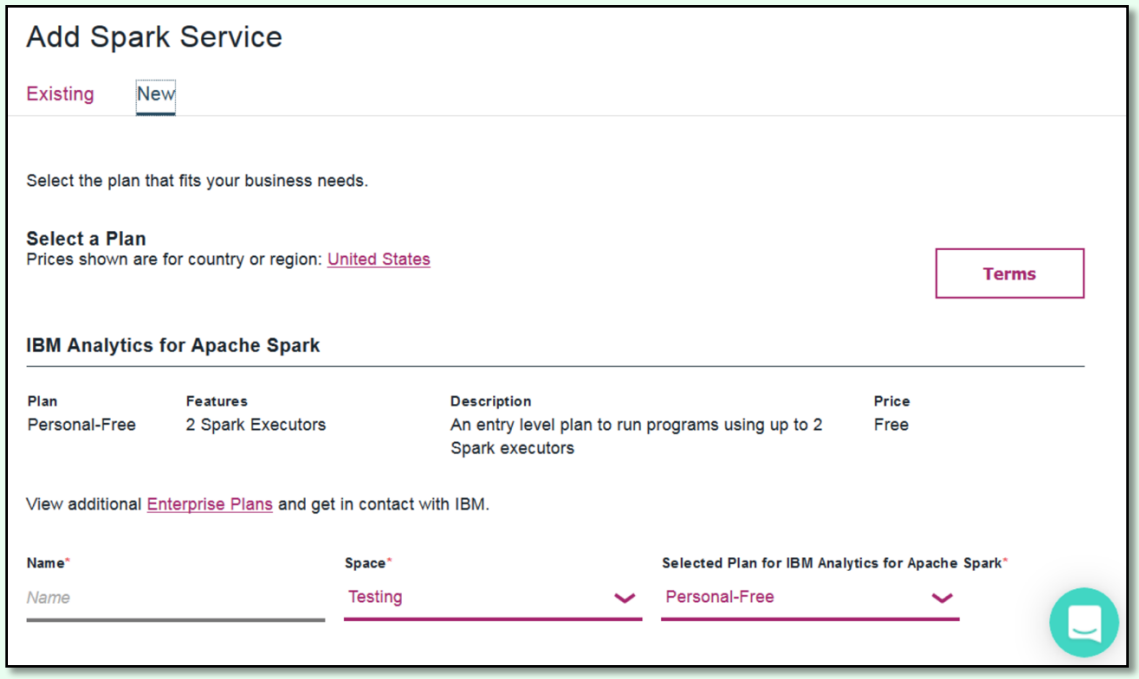
On the previous screenshot, you can see all the Projects you have, and more interesting, the Projects that others are sharing with you. You can also see the *Collaborators*, the *Creator*, and the *Last Modified* date.

You can create a new project clicking in "+ create project" button: 

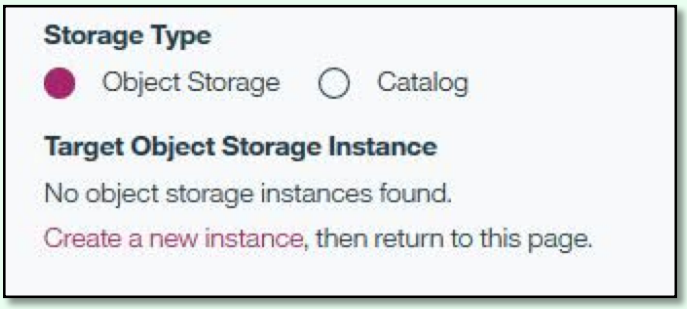
NOTE 1: If this is your first time, this message may appear: Click on the link to create a new **Apache Spark** Service on Bluemix.



If you already have a Spark service created on Bluemix, you can use it. If not, **add a new one** following the wizard that you can see on the next screenshot. For testing purposes use the **Personal-Free** plan.

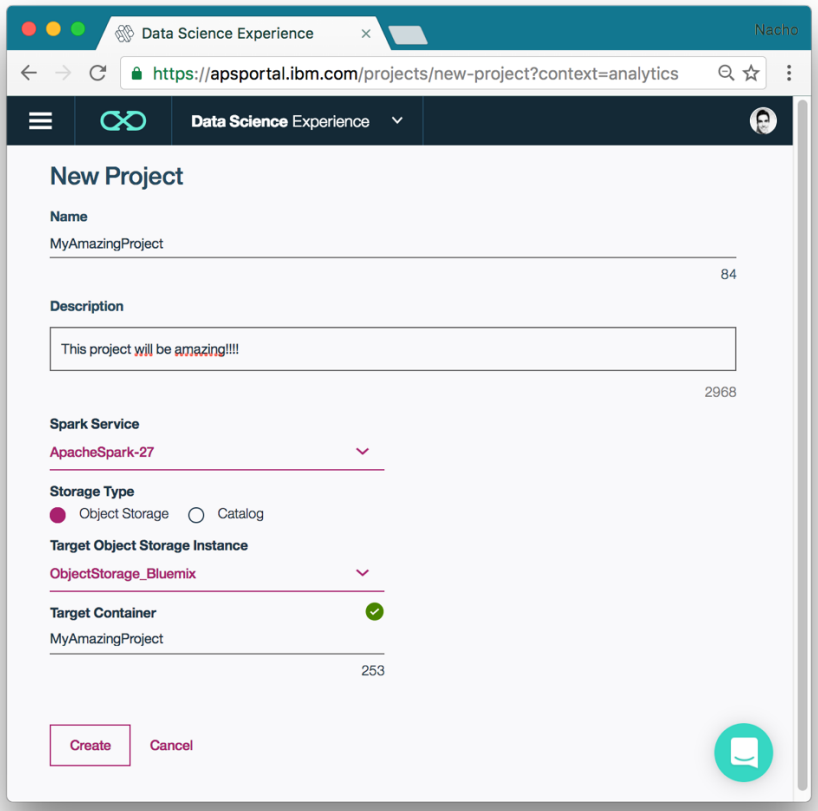


NOTE 2: Another message than may be appears (similar to the previous one), is related with the **Object Storage** service. If you don't have an Object Storage service already created in Bluemix, click on the link below "Create a new instance" to create a new one.



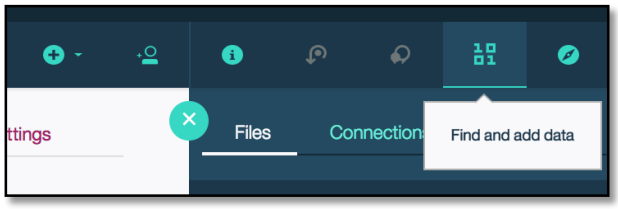
Write the Name of the Project (I'm going to use **MyAmazingProject**) and a short description. Select your Spark Service.

On the Storage Type, select Object Storage, and select the Object Storage created in Bluemix. Click on the "Create" button.

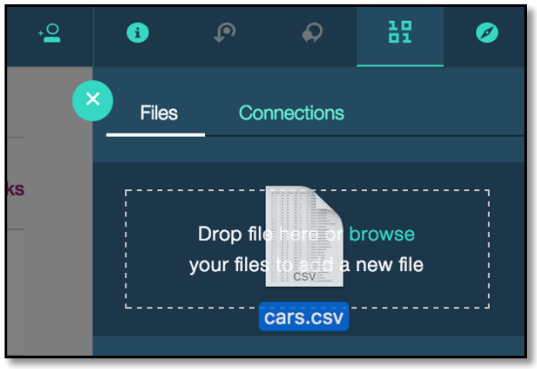

Data Science Experience
Nacho
https://apsportal.ibm.com/projects/new-project?context=analytics
Data Science Experience
New Project
Name
MyAmazingProject
84
Description
This project will be amazing!!!!
2968
Spark Service
ApacheSpark-27
Storage Type
Object Storage
Catalog
Target Object Storage Instance
ObjectStorage_Bluemix
Target Container
MyAmazingProject
253
Create
Cancel

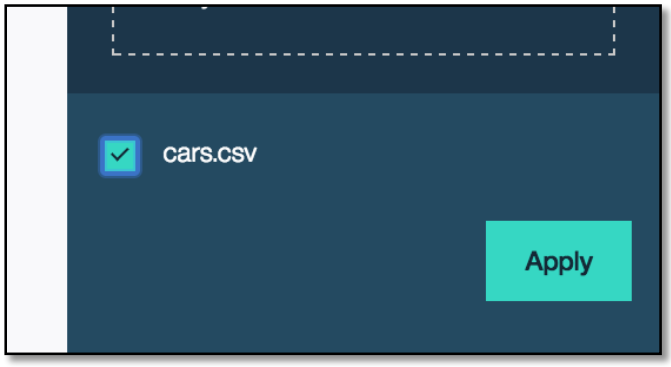
4.2. Copying a Notebook from an URL

First, we are going to upload some data. Click on the "1001 button".

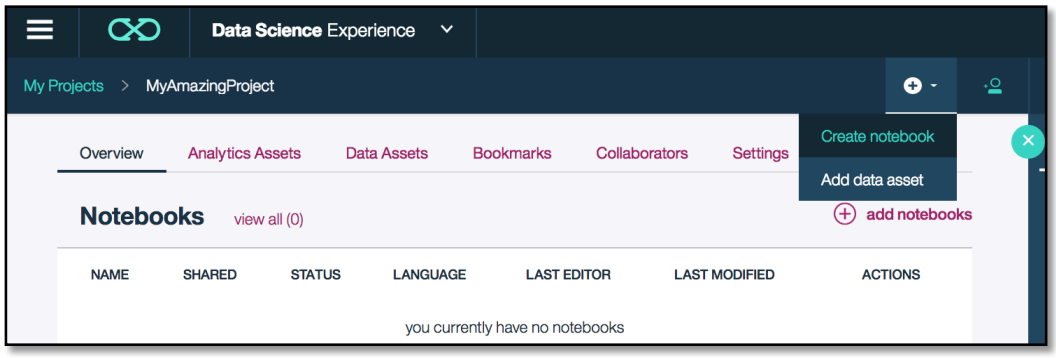


Go to the downloaded folder on your laptop and *drag&drop* the `cars.csv` file to the blue menu. Then select the csv file, and click on "Apply" button to add it to your "Data Assets":





On your new amazing project, click on the top menu "Plus" button or in the "add notebooks" button.

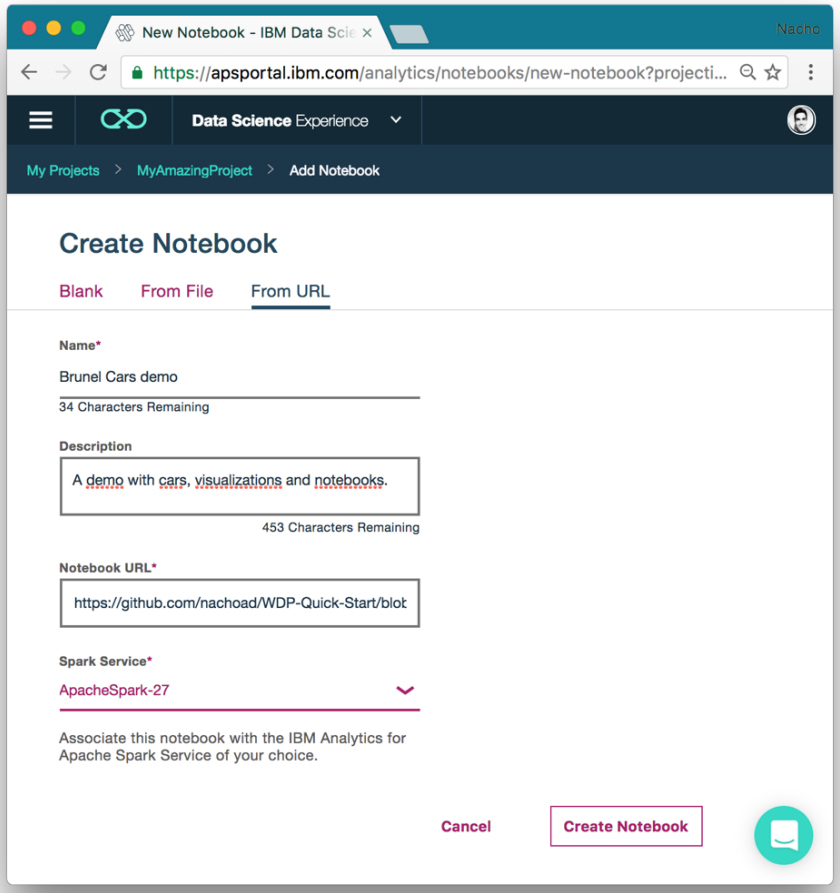


On the new wizard page, select the **`From URL`** tab.

Then give a name, for example **`Brunel Cars demo`**. Write a short description. And copy/paste this URL to the "Notebook URL" form:

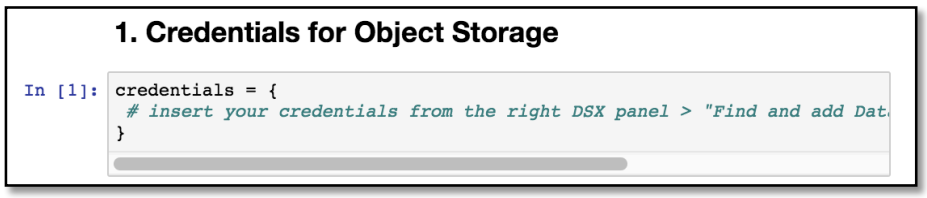
```
https://github.com/nachoad/WDP-Quick-Start/blob/master/python/Brunel-Cars-demo.ipynb
```

Then click on "Create Notebook" button.

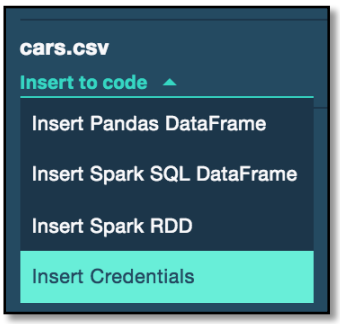


We are going to add your credentials on the notebook.

Scroll down and find this cell. Click on the cell, delete the content of the cell:



Now click on the 1001 button. You should be able to see the `cars.csv` file on the right panel. Click on Insert to code > Insert Credentials:



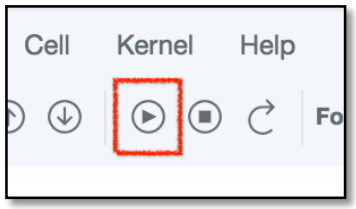
You should see a cell code similar to this:

1. Credentials for Object Storage

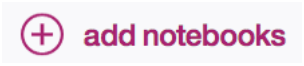
```
In [7]: # @hidden_cell
credentials = {
  'auth_url':'https://identity.open.softlayer.com',
  'project':'object_storage_52a8a3be_8a06_4188_b2fd_1b782fa33f13',
  'project_id':'daff1c142f284605bb9b6a6fcd1d4116',
  'region':'dallas',
  'user_id':'n1n1n1n1n1n1nn1',
  'domain_id':'n1n1n1n1n1n1n1n1n1',
  'domain_name':'1041753',
  'username':'nnnnnnnnnnn',
  'password':"""nnnnnnnnnnnn""",
  'container':'MyAmazingProject',
  'tenantId':'undefined',
  'filename':'cars.csv'
}
```

Verify that the variable name is "**credentials**".

Ok, now we are ready to execute the notebook and see the results. To do that, click on the play button until the end of the notebook.

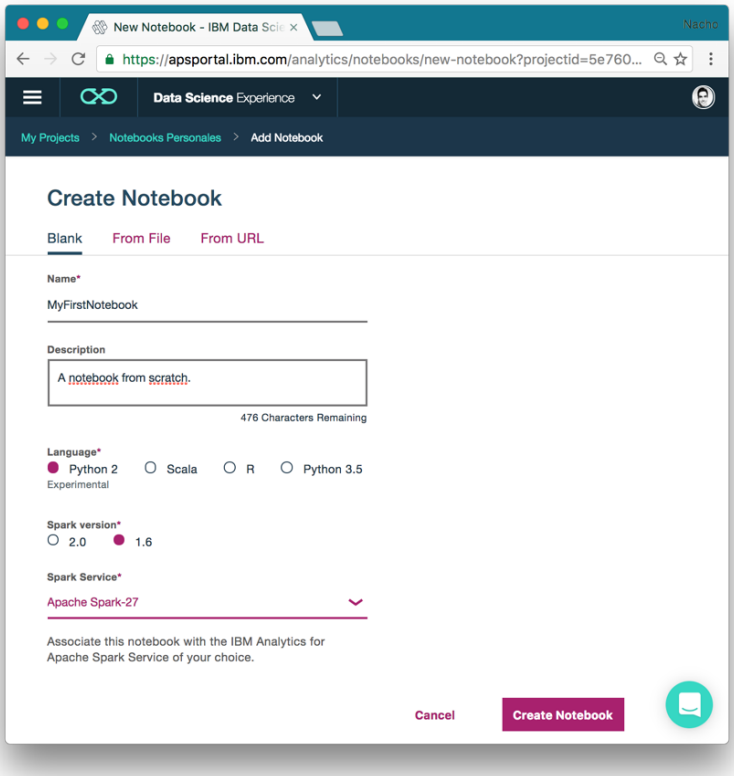


4.3. Creating my first Notebook with Python

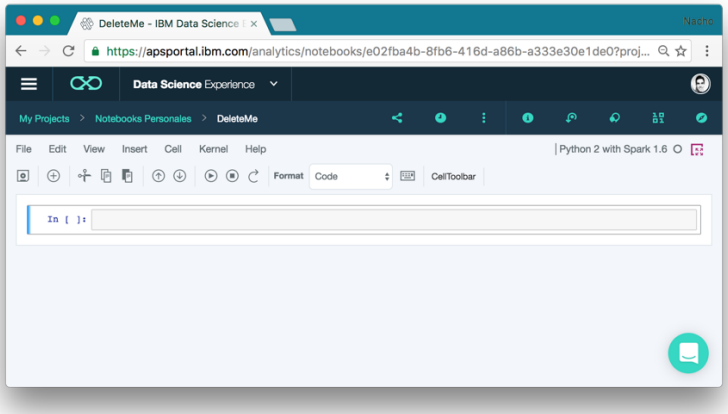
Now you've seen what a notebook looks like. Let's create a new notebook from scratch. Go to your Project a click in "Add Notebooks" button: 

Then, be sure that you have selected "Blank" as type. Write a name **MyFirstNotebook** and a short description. Select Python 2 as Language, and Spark 2.0. Your "Spark Service" should be selected, if not, select it. Click on the "Create Notebook" button.

If you see a Spark warning message saying: No Spark Service could be found for the given… please go to the `NOTE 1` on the point 4.1 of this guide.

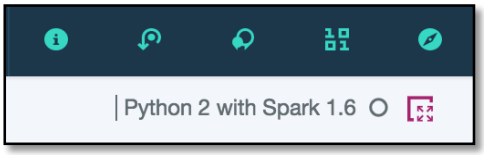


On the next screenshot, you can see the first screen you should see once you have created the new notebook. Now you are in **Edit Mode**. You know that you are in Edit Mode because you can see the upper menu (File, Edit, View… and all the icons below).

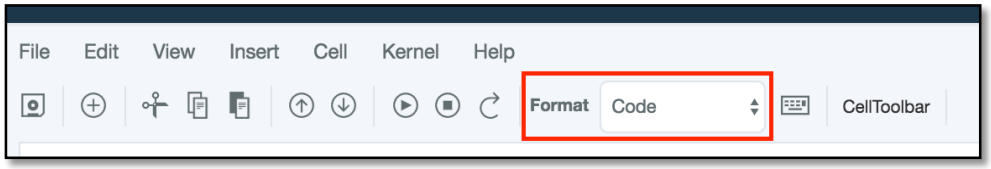


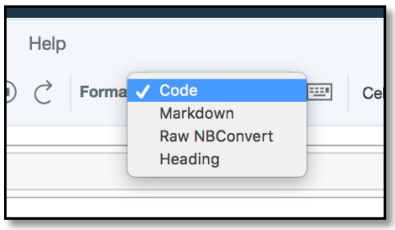
Exploring the notebook

On the upper right corner, you will see the **Language** and version of the current notebook, and the **Apache Spark version** service you have. You always can change the Language on the menu: Kernel > Change kernel.



One of the most important menus is the **Format**:

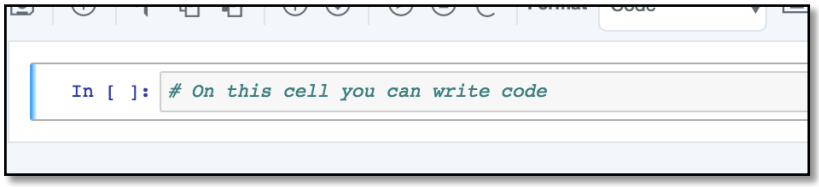




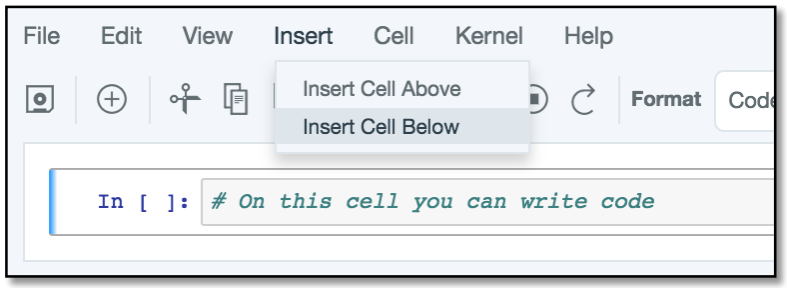
There are 4 types of formats. These are the formats for **each** cell. Usually you only use the first 2 formats:

- **Code**: A cell that contains code. If you have selected Python as a language, this cell will be interpreted as a Python Code. You always can change the Language on the menu: Kernel > Change kernel.
- **Markdown**: Markdown is a lightweight markup language with plain text formatting syntax designed so that it can be converted to HTML and many other formats using a tool by the same name. You can find a good Cheat sheet of markdown on this website: https://guides.github.com/features/mastering-markdown/

As you can see, the first cell created by default is a code cell:

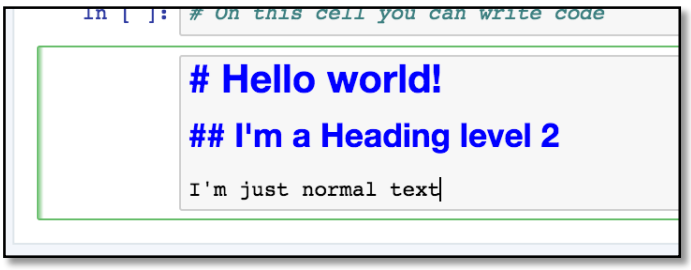


You can insert a new cell clicking on Insert > Insert Cell (Above or Below).



To change the cell format, click on the cell and then click on the Format menu to change it (for example to markdown).

Look at the cheat sheet to know how to write headings, imgs, and other stuff. For example you can write:

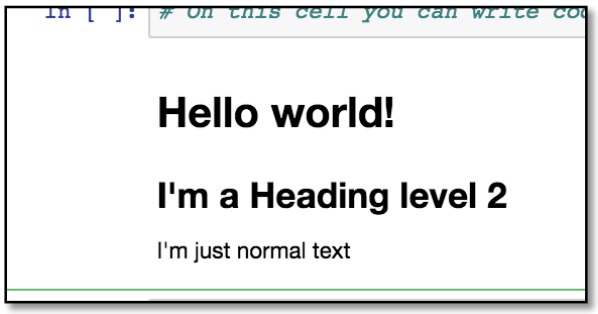


Executing the cells.

To execute a cell, click on the cell, and click on the play button:

Another easy way to execute is click on the cell and on your keyboard press Shift + Enter.

If you execute the last markdown, you should see that result:



We will use **GitHub** again, to see the notebook you should obtain. Open a new tab on your web browser, and go to this URL: https://github.com/nachoad/WDP-Quick-Start/blob/master/python/MyFirstNotebook.ipynb

Try to replicate this notebook on yours.

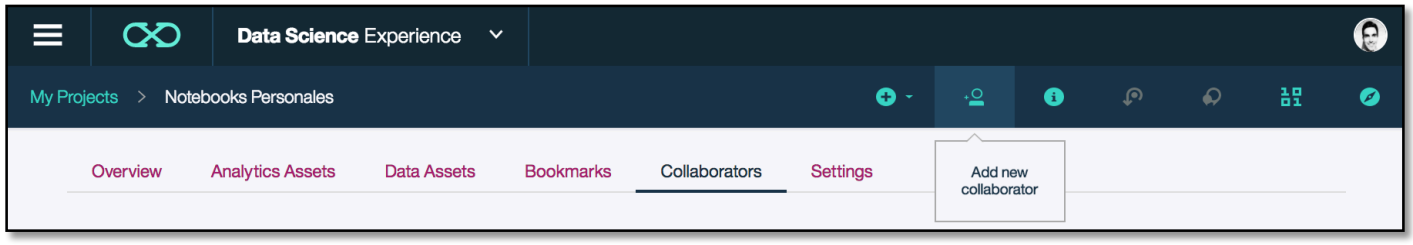
If you have doubts, you can follow this file with a plain text: https://github.com/nachoad/WDP-Quick-Start/blob/master/python/MyFirstNotebook-plain.py

4.4. Sharing the Notebook internally & externally

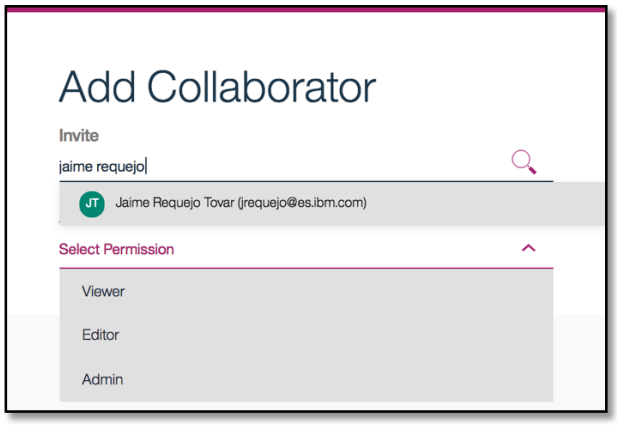
Now we will see how you can share your notebook.

Internally

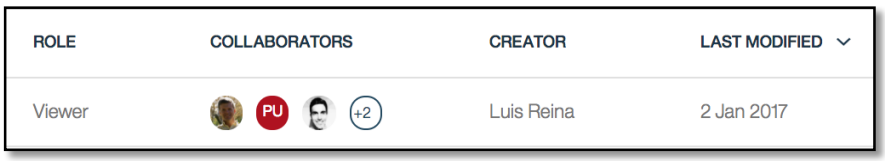
First go to your project, and on the top menu click on the "Add new collaborator" button (or click on Collaborators tab > add collaborators).



On the wizard, you can search by *name* or *email*, and select the kind of permission (Viewer, Editor or Admin) to share the project:



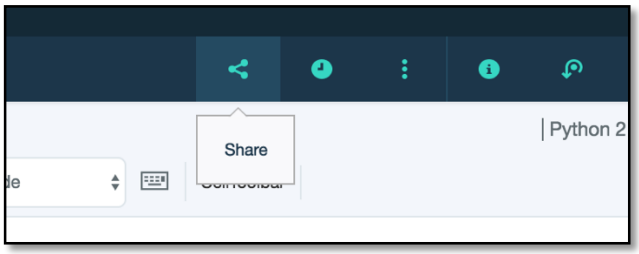
This way you can have a team of people working on the same project, and with different roles:

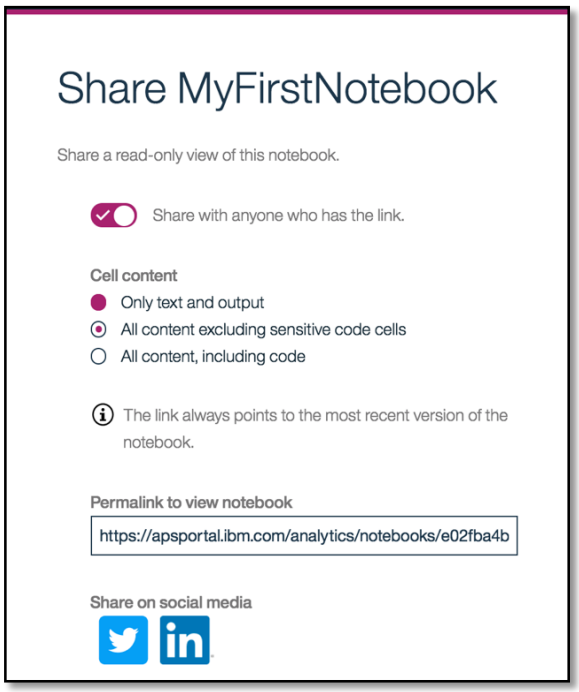


Externally

Now you are going to share it with someone who is **external** of your company/department. We are going to generate one URL with the notebook link.

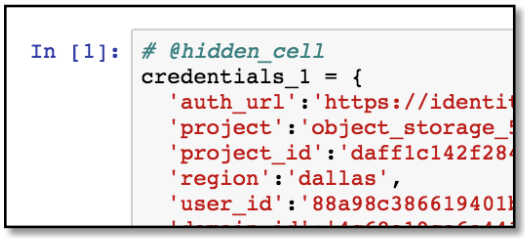
Open the notebook you want to share. Click on the "Share" button:





As you can see, you can select how to protect some confidential cells.

For example, you can share the entire notebook, but protect the sensitive code cells that have passwords by selecting the second option in "Cell content". This will share all the notebook cells, excluding the cells with the tag `# @hidden_cell` on the code:

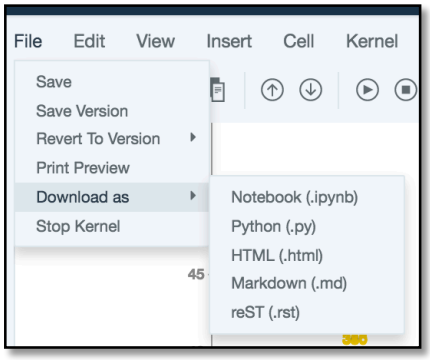


Or you can just share the text and output cells.

This wizard generates a Permalink to view your notebook, that you can send to anybody by email, or share in **Twitter** or **LinkedIn**.

Download a local copy

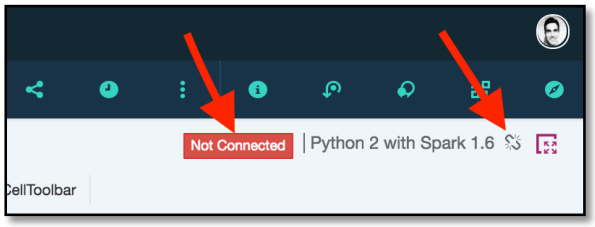
In DSX, you are able also to download a local copy to your laptop. Go to one of your notebooks, and click on File > Download as >.


File
Edit
View
Insert
Cell
Kernel
Save
Save Version
Revert To Version
Print Preview
Download as
Stop Kernel
Notebook (.ipynb)
Python (.py)
HTML (.html)
Markdown (.md)
reST (.rst)
45

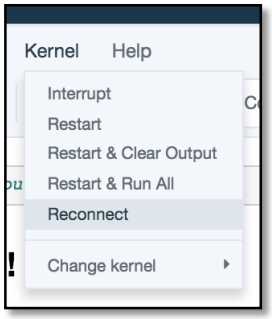
5. FAQs & Documentation

Kernel not Connected

Issue: If you stay a long time without work with your notebook, the Kernel stops and the notebooks disconnects. This is a normal action on notebooks.



Solution: You can reconnect on the menu: Kernel > Reconnect.

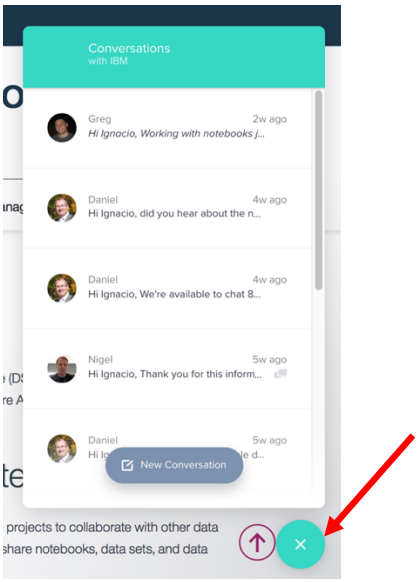


Official Documentation

Here there is a link to the FAQs and the Documentation of the Platform:

http://datascience.ibm.com/docs/content/getting-started/faq.html

On the lower right corner, there is always a bubble of help available. Click on it to open a chat with our support team.



6. Reviews and versions

Faith of errors. If you find something erroneous on this Quick Start guide, please feel free to send an email to ignacio.alonso@es.ibm.com with your comments.

Or go to the GitHub repository and create a new Issue: https://github.com/nachoad/WDP-Quick-Start/issues .

Thank you very much for your help.

Date	What	Comments
30 December 2016	v.1.0 of this document.	First release.
9 January 2016	v.1.1 of this document	-
10 January 2016	v.1.2 of this document	Note 2 added to the document. Message that appears if there is not an Object Storage already created.

Author: Ignacio Alonso Delgado