

# **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

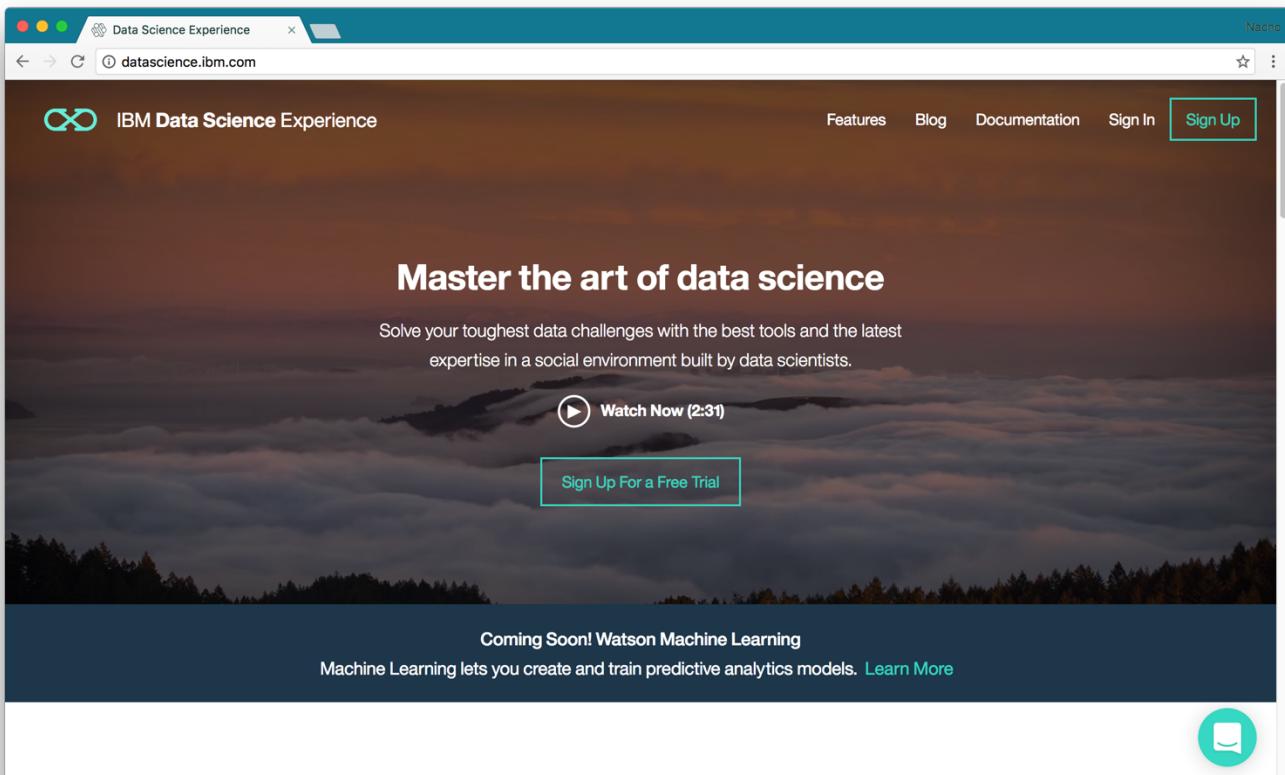
1. REGISTRATION ON DATA SCIENCE EXPERIENCE (DSX)	3
1.1. FIRST TIME USING BLUEMIX	3
1.2. I ALREADY HAVE A BLUEMIX ACCOUNT	5
2. DISCOVERING DSX – DATA HUB	6
3. CREATING A NEW DATA HUB CATALOG	8
3.1. UPLOADING DATA TO THE CATALOG	10
3.2. SHARING THE CATALOG WITH COLLEAGUES	12
4. DISCOVERING DATA SCIENCE EXPERIENCE (DSX)	13
4.1. CREATING A NEW PROJECT	14
4.2. COPYING A NOTEBOOK FROM AN URL	18
4.3. CREATING MY FIRST NOTEBOOK WITH PYTHON <i>Exploring the notebook</i>	21
<i>Executing the cells.</i>	23
4.4. SHARING THE NOTEBOOK INTERNALLY & EXTERNALLY <i>Internally</i>	24
<i>Externally</i>	25
<i>Download a local copy</i>	25
5. FAQS & DOCUMENTATION	27
<i>Kernel not Connected</i>	27
<i>Official Documentation</i>	27
6. REVIEWS AND VERSIONS	28

# 1. Registration on Data Science Experience (DSX)

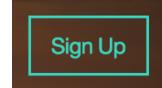
Enter on a web browser. (It's recommended to use Google Chrome).

Type the following URL: [datascience.ibm.com](http://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.

The screenshot shows the 'Data Science Experience' registration page at [datascience.ibm.com/registration/stepone](https://datascience.ibm.com/registration/stepone). The page has a dark blue header with the IBM logo and navigation links for 'Sign In' and 'Sign Up'. The main content area features a heading 'Try a preview version of Data Science Experience powered by Bluemix.' Below it, there's a section about the Apache Spark plan ('Personal') and Object Storage plan ('Free'). A red box highlights the 'Create your Bluemix Account:' input field where the email 'myemail@mycompany.com' is entered. To the right of the input field is a 'Continue' button. Below the input field, there are links for existing users ('Already have a Bluemix account?') and for continuing with credentials ('Continue with your Bluemix Credentials'). A green message bubble icon is visible on the right side.

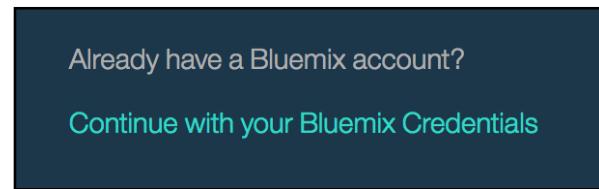
On the next screen, enter your info. And check your email (entered on the previous screen) to copy the **Confirmation Code**:

The screenshot shows the 'Register for Bluemix' form at [https://console.ng.bluemix.net/registration/link?email=myemail%40mycompany.com&success\\_uri=https%3A%2F%2Fapsportal.ibm.com%2Fr...](https://console.ng.bluemix.net/registration/link?email=myemail%40mycompany.com&success_uri=https%3A%2F%2Fapsportal.ibm.com%2Fr...). The form includes fields for First Name (Ignacio), Last Name (Alonso Delgado), Phone Number (555 555 555), Country or Region (Spain), Email (myemail@mycompany.com), Password, and Re-enter Password. A red box highlights the 'Email' field. A red arrow points from the text 'The Confirmation Code should be in your email inbox' to the 'Email' field. A green message bubble icon is visible on the right side.

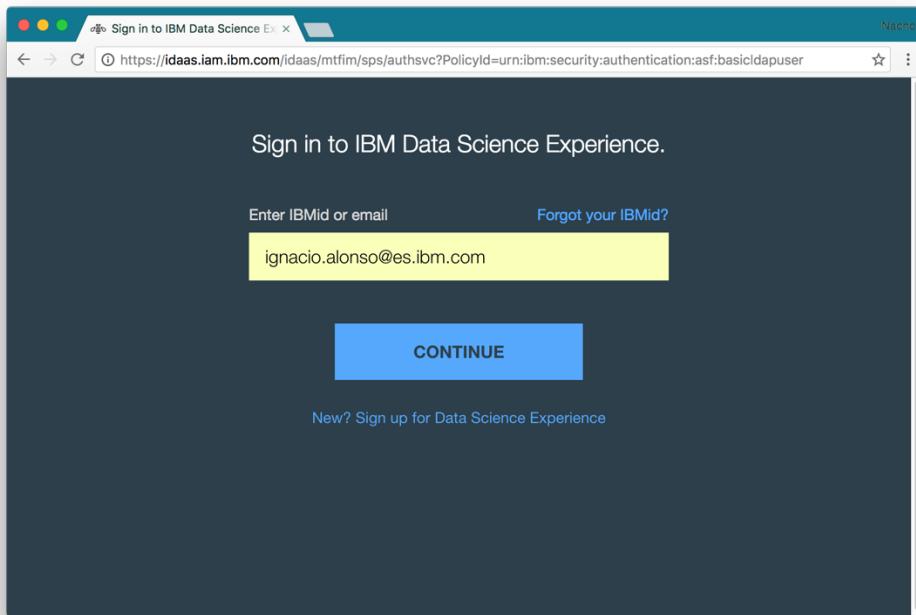
Then, create a new Organization and Space in Bluemix.

## 1.2. I already have a Bluemix Account

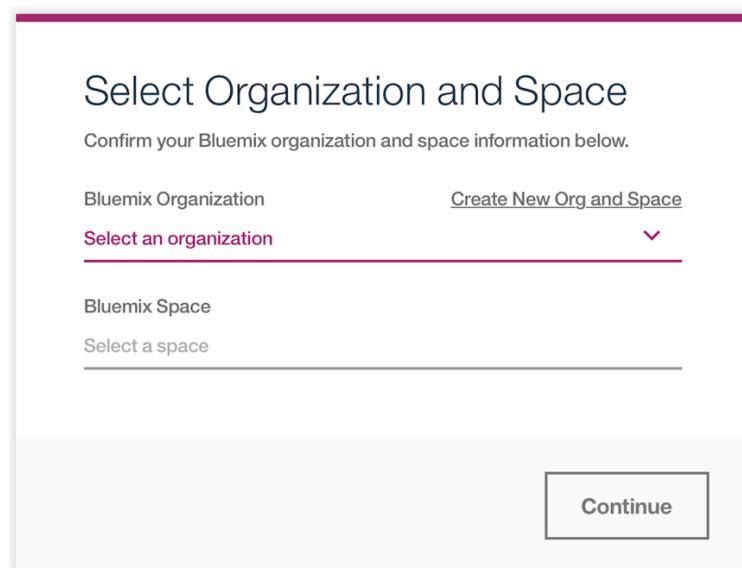
If you already have a Bluemix account, just click on "Continue with your Bluemix Credentials":



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

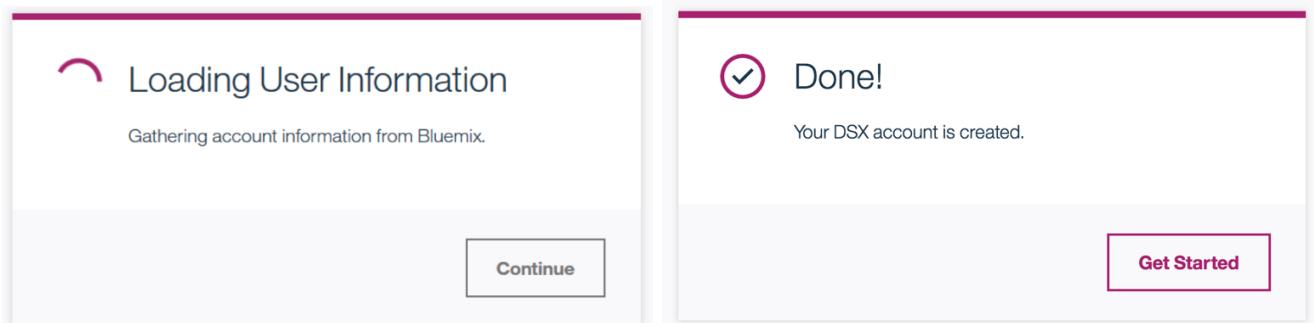


If the next pop-up window appears, select your Bluemix Organization and Space. Click on Continue:



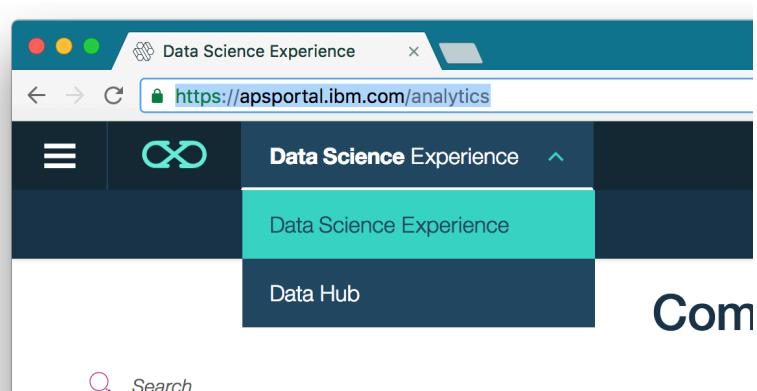
## 2. Discovering DSX – Data Hub

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:

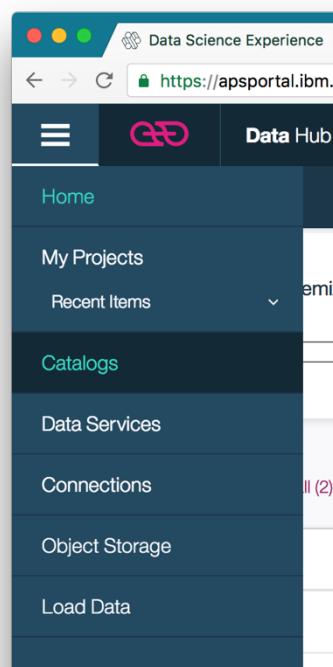
- **Object Storage:** You have 5GB of Object Storage space for free in the free account.
- **Catalogs:** A catalog is a cloud-based data repository for structured and unstructured data from a variety of sources. You copy and categorize data assets into the catalog and add collaborators who can analyze 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.

The screenshot shows the Data Science Experience interface. At the top, there's a header with a logo, a search bar containing 'https://apsportal.ibm.com/overview', and a user profile icon. Below the header, the main content area is divided into several sections:

- Object Storage**: Shows a single entry: 'ObjectStorage\_Bluemix' with '355.64 kB used' and '5.00 GB free'.
- Catalogs**: Shows two entries: 'HRCatalog' (Admin, 30 Dec 2016) and 'Catologo Luis' (Viewer, 16 Dec 2016). There's a 'create catalog' button.
- Projects**: Shows one entry: 'Notebooks'. There's a 'create project' button.

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.

NAME ^	ROLE	COLLABORATORS	CREATOR	LAST MODIFIED	ACTIONS
Catologo Luis	Viewer		Luis Reina	00:18 PM UTC, 2016/12/16	...
HRCatalog	Admin		Ignacio Alonso	01:08 PM UTC, 2016/12/30	...

Click on “+ New Catalog” button Write a new name (for example TrainingCatalog) and a short description. Click on “Create”

New Catalog

Name  
TrainingCatalog

Description  
This Catalog will be used for training purposes.

Create Cancel

TrainingCatalog

Upload Browse Access Control Dashboard

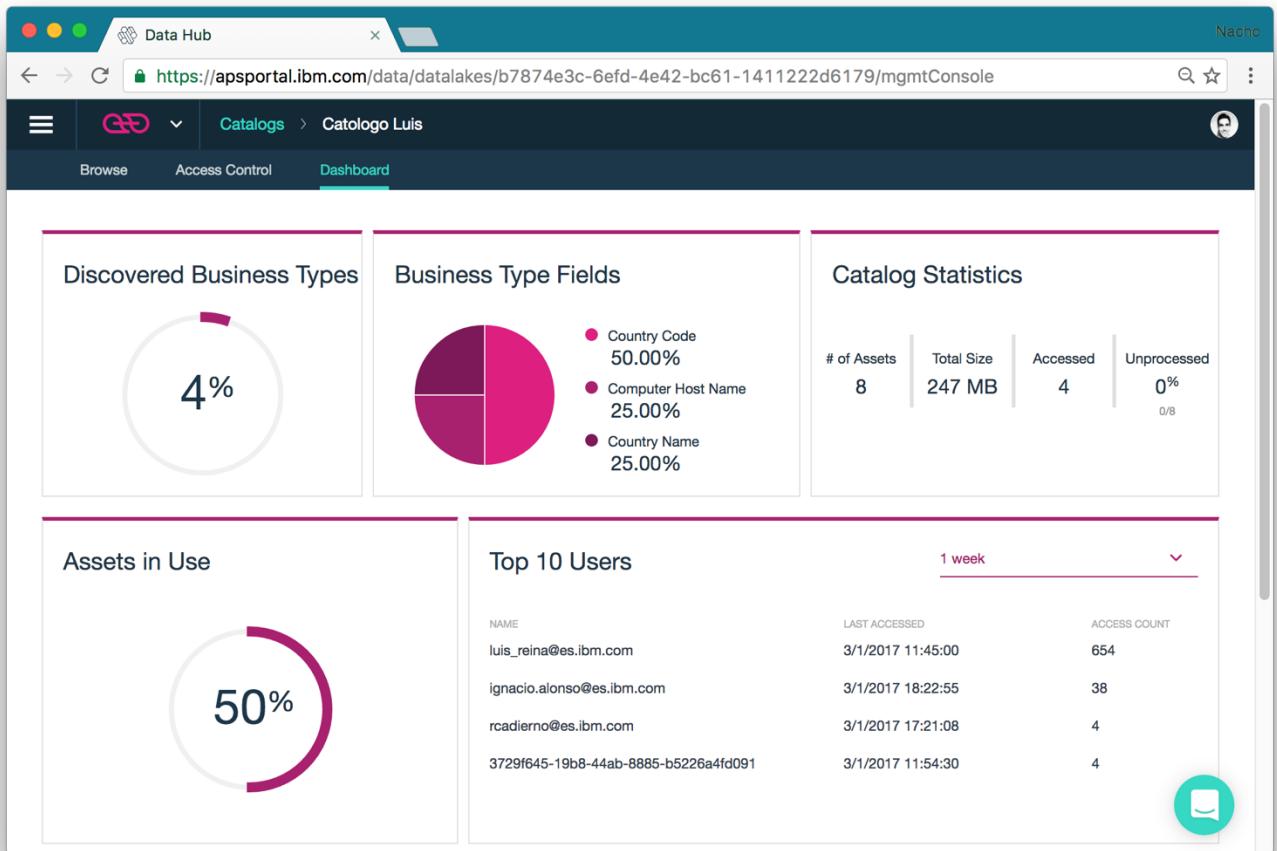
Search asset names or tags

Tags No tags are defined

Recently Accessed Data Sets 0 of 0 assets

Create Data Set Start

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.



## 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.

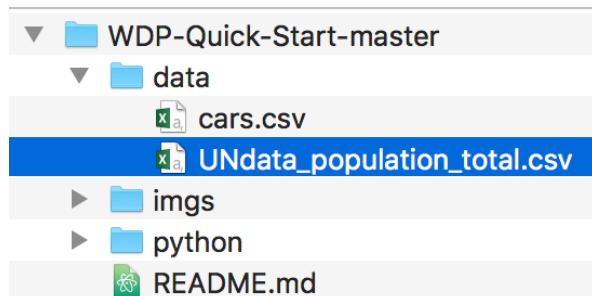
We will use this **GitHub** repository: <https://github.com/nachoad/WDP-Quick-Start>

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

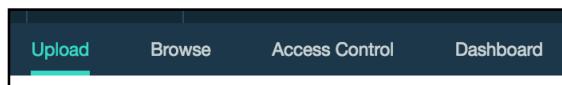


<https://github.com/nachoad/WDP-Quick-Start/archive/master.zip>

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

[Upload](#)

select the Origin country: US. Click on Upload button.



Where is the data you want to upload?

Local File	Stored Connection (Ex. dashDB or DB2 on Cloud)
------------	---

Identifying your source and target

Source

Browse	UNdata_population_total.csv	X
--------	-----------------------------	---

Target

Destination

SHARED ZONE

Name

population

90

Description

The World Development Indicators (WDI) is the statistical benchmark that helps measure the progress of development.

185

Tags

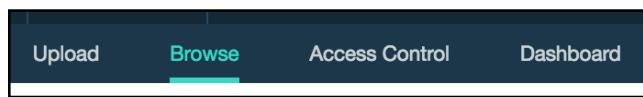
population countries

80

Origin country

US

Go to the Browse tab to see the data uploaded.



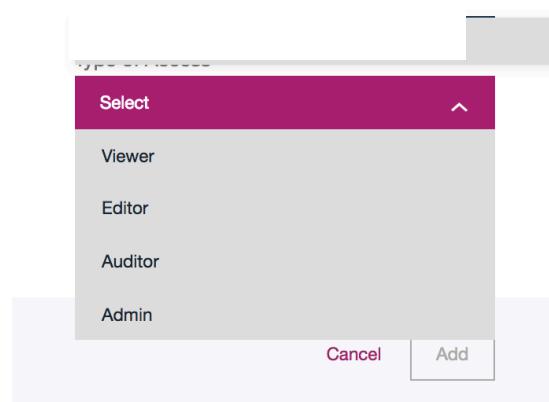
## 3.2. Sharing the Catalog with Colleagues

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

The screenshot shows a web browser window titled "Data Hub" with the URL <https://apsportal.ibm.com/data/datalakes/95bf792e-a39...>. The page is titled "Catalogs > TrainingCatalog". The "Access Control" tab is selected. A table lists a single collaborator: Ignacio Alonso (ignacio) with Admin permission. There is a search bar labeled "Find in Collaborators" and a teal "add collaborators" button with a plus sign.

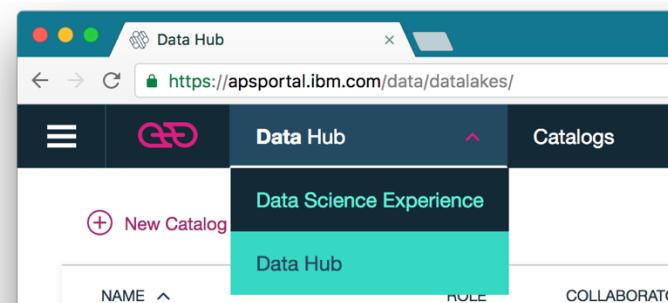
### Add New Collaborator

Add users to your catalog for collaboration.



## 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, ...

The screenshot shows the Data Science Experience Community page. At the top, there's a search bar and navigation tabs for 'All', 'Articles', 'Data Sets', 'Notebooks', and 'Tutorials'. Below these are sections for 'Articles' and 'Data Sets', each displaying four items with thumbnails, titles, authors, dates, and formats.

Category	Title	Author	Date	Format
Articles	R Markdown Reference Guide	RStudio	Dec 29, 2016	Web page
Articles	On calculating AUC	Win-Vector Blog	Dec 29, 2016	Web page
Articles	Data Wrangling with dplyr and tidyR Cheat...	RStudio	Dec 28, 2016	Web page
Articles	Hyperparameter Optimization: Sven Hafener	Spark.tc	Dec 28, 2016	Video
Data Sets	Airbnb Data for Analytics: Vienna Reviews	IBM	Dec 20, 2016	
Data Sets	GoSales Transactions for Naive Bayes Model	IBM	Dec 20, 2016	
Data Sets	Airbnb Data for Analytics: Washington D.C....	IBM	Dec 20, 2016	
Data Sets	Airbnb Data for Analytics: Washington D.C....	IBM	Dec 20, 2016	

## 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.

The first screenshot shows the 'Community' section of the Data Science Experience interface. The second screenshot shows the 'My Projects' section, where a new project named 'Default Project' has been created. The table below summarizes the project details.

Name	Role	Collaborators	Creator
Default Project	Admin		Ignacio Alonso

NAME	ROLE	COLLABORATORS	CREATOR	LAST MODIFIED	ACTIONS
Demos	Admin	PD	Ignacio Alonso	20 Dec 2016	...
Notebooks Personales	Admin		Ignacio Alonso	17 Dec 2016	...
Analitica de Datos de Twitter	Viewer	PU  +2	Luis Reina	16 Dec 2016	...
NBA	Admin		Ignacio Alonso	10 Dec 2016	...
Proyecto por defecto	Admin		Luis Reina	22 Nov 2016	...
Pruebas con Scala	Admin		Ignacio Alonso	6 Oct 2016	...

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.

**Spark Service\***

No Spark Services could be found for the given project.  
Please select another project or add a Spark Service to the project [here](#). Then return to this page to create a new notebook.

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.

**Add Spark Service**

[Existing](#) [New](#)

Select the plan that fits your business needs.

**Select a Plan**  
Prices shown are for country or region: [United States](#)

[Terms](#)

**IBM Analytics for Apache Spark**

Plan	Features	Description	Price
Personal-Free	2 Spark Executors	An entry level plan to run programs using up to 2 Spark executors	Free

[View additional Enterprise Plans](#) and get in contact with IBM.

**Name\***  **Space\***  **Selected Plan for IBM Analytics for Apache Spark\***



**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.

**Storage Type**

Object Storage  Catalog

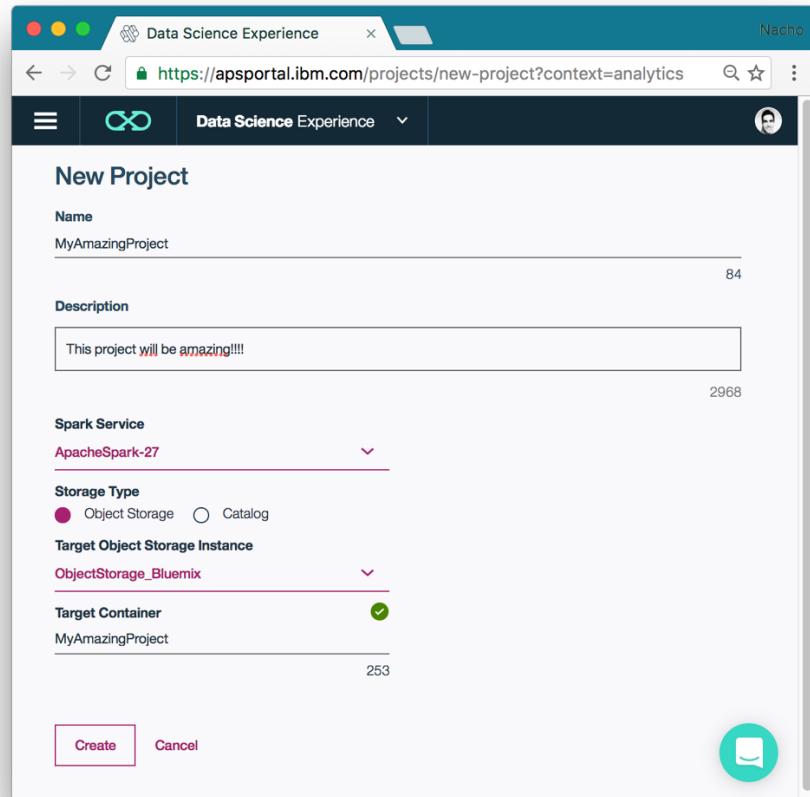
**Target Object Storage Instance**

No object storage instances found.

[Create a new instance](#), then return to this page.

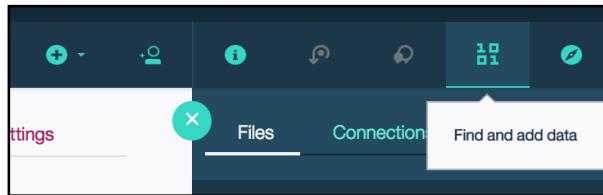
Write the Name of the Project (I'm going to use **MyAmazingProject**) and a short description. Select your Spark Service.

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

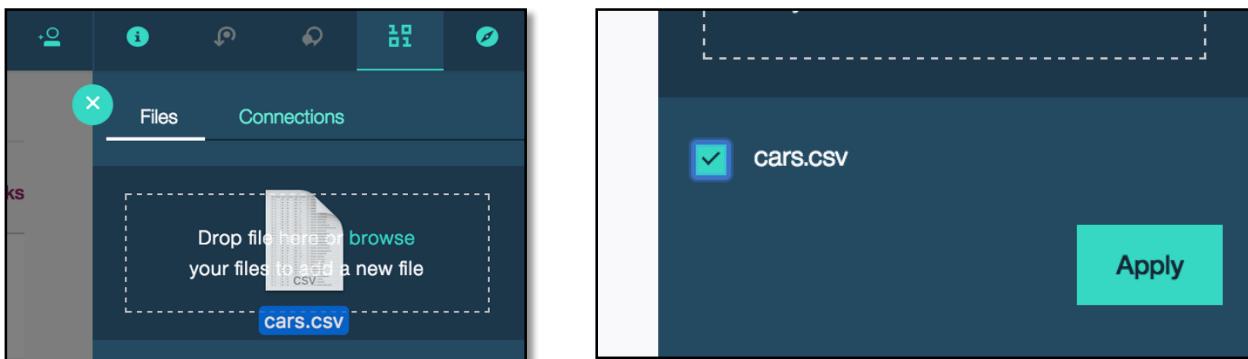


## 4.2. Copying a Notebook from an URL

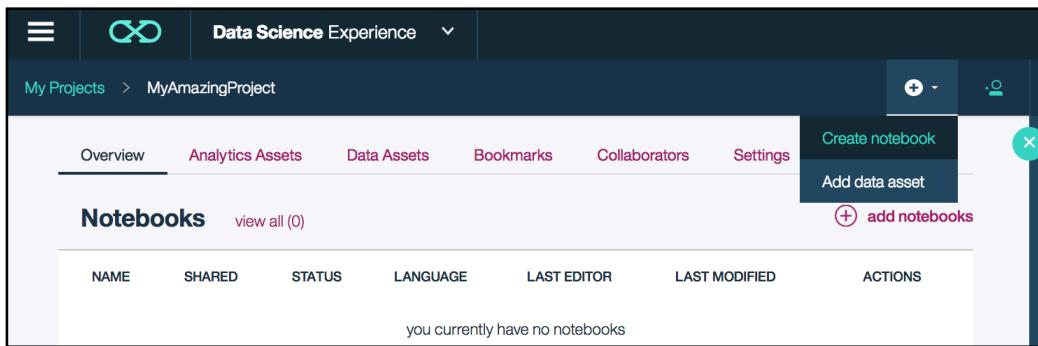
First, we are going to upload some data. Click on the “1001 button”.



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.

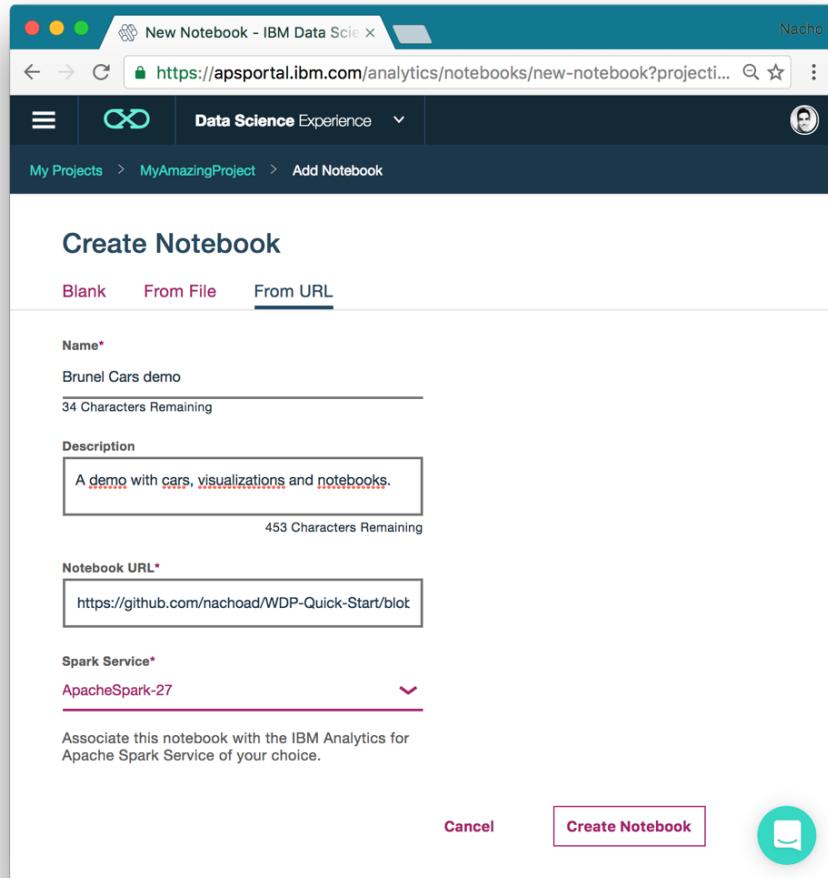


On the new wizard page, select the **From URL** tab.

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.



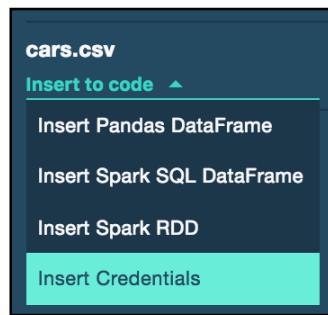
We are going to add your credentials on the notebook.

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

### 1. Credentials for Object Storage

```
In [1]: credentials = {
    # insert your credentials from the right DSX panel > "Find and add Data Source"
}
```

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:



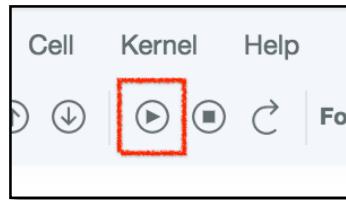
You should see a cell code similar to this:

## 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': 'dafffc142f284605bb9b6a6fc1d4116',
    'region': 'dallas',
    'user_id': 'nnnnnnnnnnnnnnnn',
    'domain_id': 'nnnnnnnnnnnnnnnn',
    'domain_name': '1041753',
    'username': 'nnnnnnnnnnnn',
    '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.

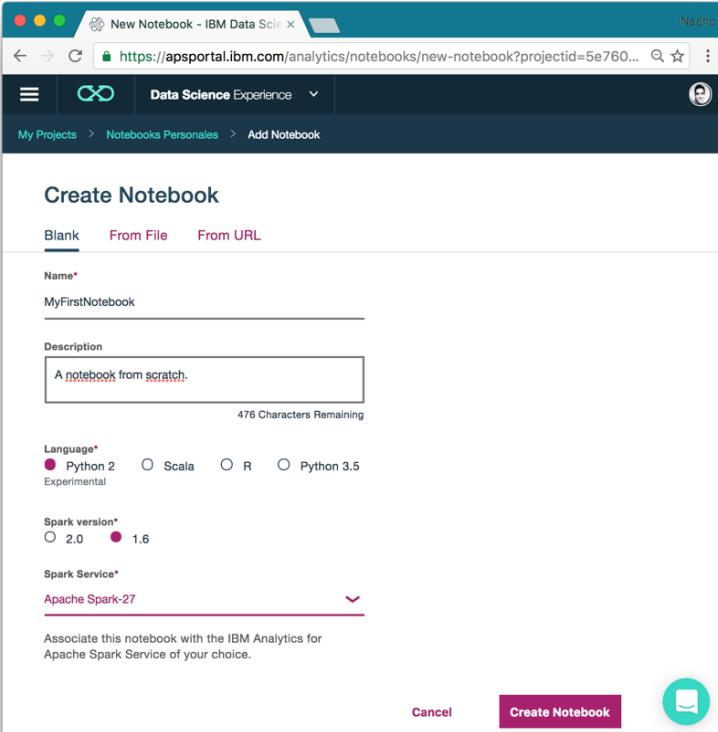


## 4.3. Creating my first Notebook with Python

Now you've seen what a notebook looks like. Let's create a new notebook from scratch. Go to your Project

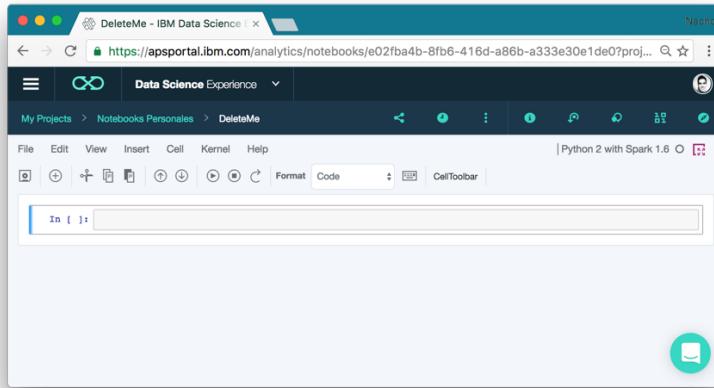
a click in "Add Notebooks" button:  **add notebooks** Then, be sure that you have selected "Blank" as type. Write a name **MyFirstNotebook** and a short description. Select Python 2 as Language, and Spark 2.0. Your "Spark Service" should be selected, if not, select it. Click on the "Create Notebook"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



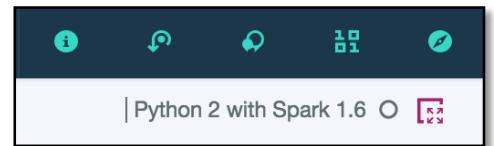
The screenshot shows the 'Create Notebook' dialog box. At the top, there are three tabs: 'Blank' (selected), 'From File', and 'From URL'. Below the tabs, the 'Name\*' field contains 'MyFirstNotebook'. The 'Description' field contains 'A notebook from scratch.' A note below it says '476 Characters Remaining'. Under 'Language\*', 'Python 2' is selected. Under 'Spark version\*', '1.6' is selected. Under 'Spark Service\*', 'Apache Spark-27' is selected. At the bottom, there is a note: 'Associate this notebook with the IBM Analytics for Apache Spark Service of your choice.' Two buttons are at the bottom right: 'Cancel' and a purple 'Create Notebook' button with a white icon.

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).

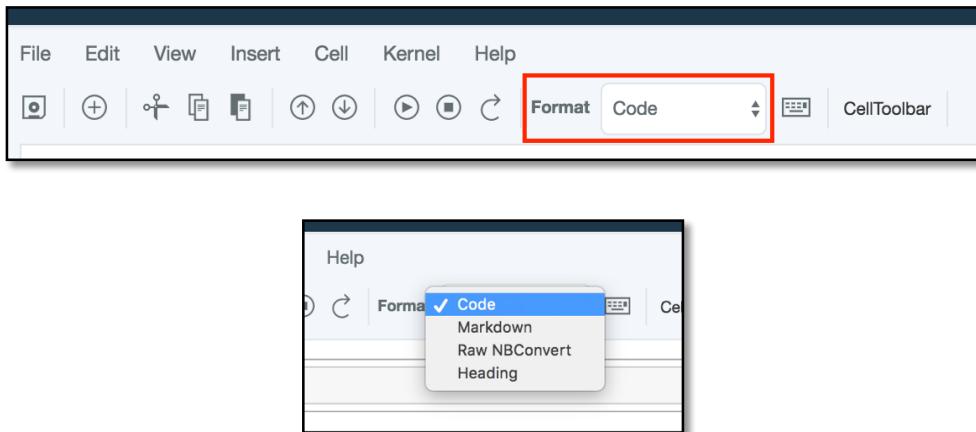


## Exploring the notebook

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.



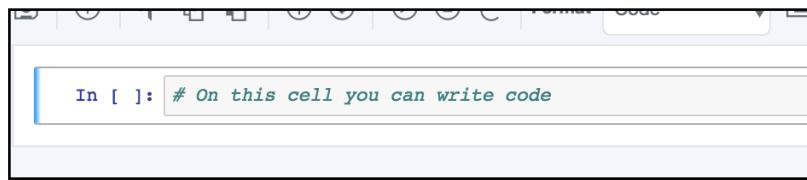
One of the most important menus is the **Format**:



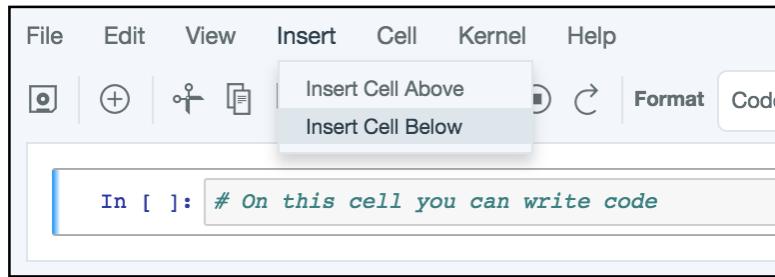
There are 4 types of formats. These are the formats for **each** cell. Usually you only use the first 2 formats:

- **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:

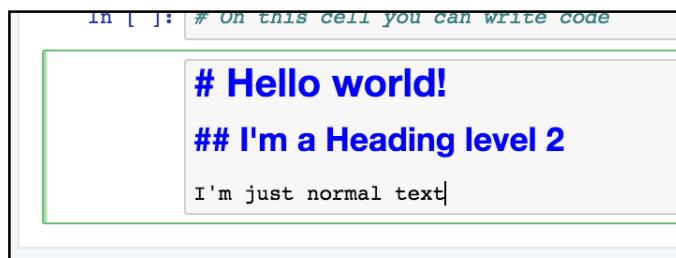


You can insert a new cell clicking on Insert > Insert Cell (Above or Below).



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:

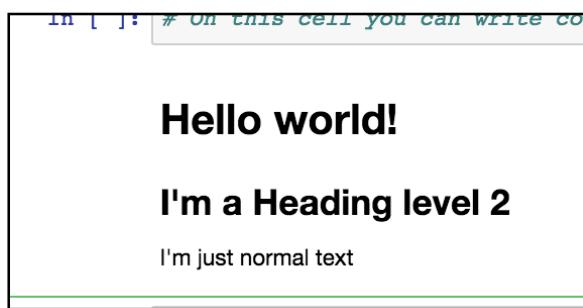


## Executing the cells.

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>

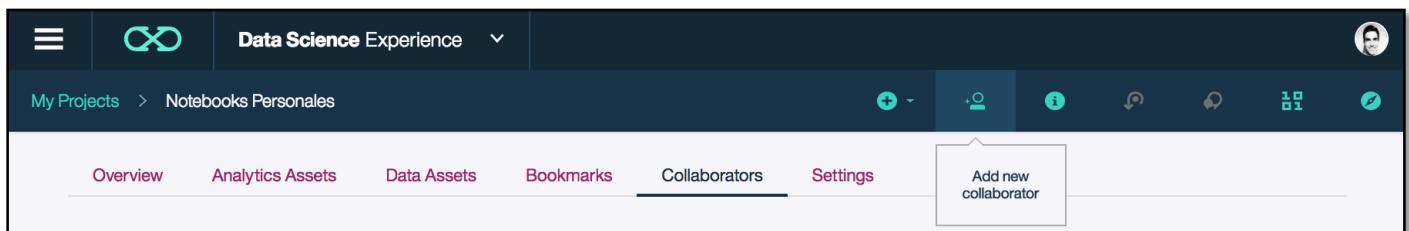
---

## 4.4. Sharing the Notebook internally & externally

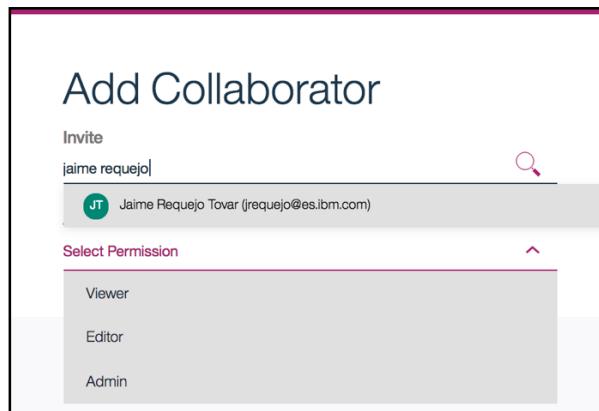
Now we will see how you can share your notebook.

### 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:

ROLE	COLLABORATORS	CREATOR	LAST MODIFIED
Viewer	+2	Luis Reina	2 Jan 2017

## 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:

The screenshot shows the Jupyter Notebook interface. In the top toolbar, there is a 'Share' button. To the right of the toolbar, a modal window titled 'Share MyFirstNotebook' is open. The modal contains the following information:

- A checkbox labeled 'Share with anyone who has the link.' which is checked.
- A section titled 'Cell content' with three options:
  - 'Only text and output' (radio button)
  - 'All content excluding sensitive code cells' (radio button, selected)
  - 'All content, including code' (radio button)
- A note: '(i) The link always points to the most recent version of the notebook.'
- A 'Permalink to view notebook' field containing the URL: <https://apsportal.ibm.com/analytics/notebooks/e02fba4b>
- A 'Share on social media' section with icons for Twitter and LinkedIn.

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:

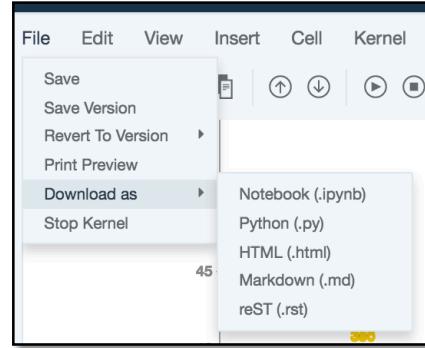
```
In [1]: # @hidden_cell
credentials_1 = {
    'auth_url': 'https://identity/api',
    'project': 'object_storage',
    'project_id': 'daff1c142f284',
    'region': 'dallas',
    'user_id': '88a98c3866194011'
}
```

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

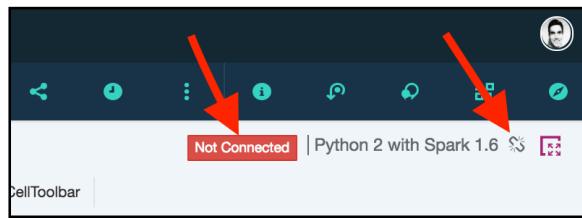
In DSX, you are able also to download a local copy to your laptop. Go to one of your notebooks, and click on File > Download as >.



## 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.

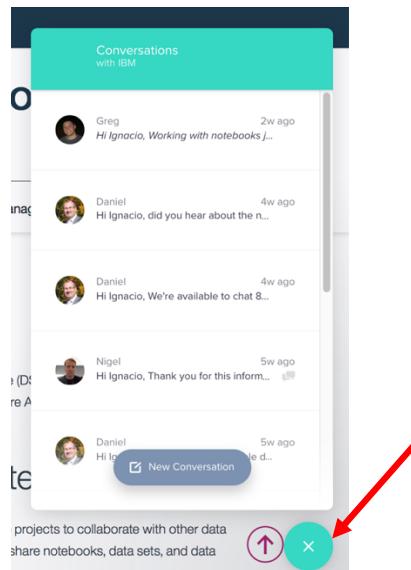


### Official Documentation

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](mailto: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
<b>30 December 2016</b>	v.1.0 of this document.	First release.
<b>9 January 2016</b>	v.1.1 of this document	-
<b>10 January 2016</b>	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**