Statistica e gestione delle informazioni

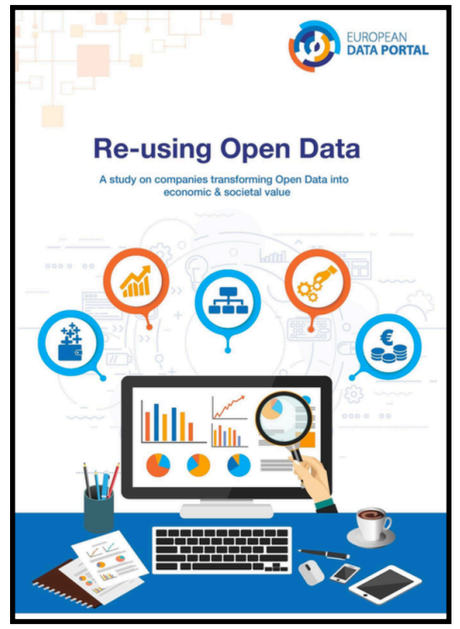
Cloud Data Science Lab

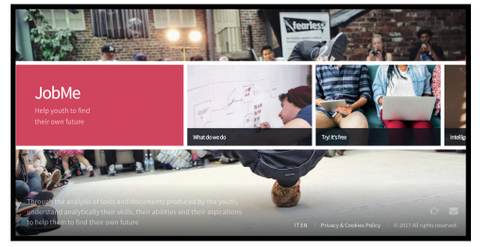
Google Colaboratory and Databricks

mauro.pelucchi@unimib.it
Mauro Pelucchi

Conosciamoci

- Mauro Pelucchi mauro.pelucchi@unimib.it mauro.pelucchi@gmail.com
- Laureato in **Ingegneria Informatica** presso Università degli Studi di Bergamo con una tesi sull'utilizzo degli OpenData (www.hammer-project.com)
- Master in **Business Intelligence & Big Data Analytics** (2015/2016, trovate il lavoro finale all'indirizzo web www.job-me.it)
- Mi occupo di **Data Science** e **Data Engineering**: progettazione e messa in qualità di grandi basi di dati, AI, machine learning, estrazione di patterns da dati non strutturati, realizzazione di sistemi direzionali statistici e strumenti per la data visualization





- In passato ho progettato e realizzato sistemi di Process Integration e Business Intelligence per grandi realtà industriali e per la pubblica amministrazione
- Main interests: **OpenData e Open Government, Big Data Analysis, Data Presentation techniques, Artificial Intelligence, Machine Learning Modeling & Deep Learning**

Cosa vedremo?

- Google Colaboratory
- DataBricks

Google Colaboratory
DataBricks

Cloud based run time python, R, scala of notebooks

Google Colaboratory
DataBricks

- Write and execute code in Python, R or Scala
- Document your code that supports mathematical equations
- Create/Upload/Share notebooks
- Import/Save notebooks from/to Google Drive or others
- Import/Publish notebooks from GitHub
- Import external datasets e.g. from Kaggle
- Integrate PyTorch, TensorFlow, Keras, OpenCV
- ~Free Cloud service with free GPU

Google Colaboratory

What is Google Colaboratory?

- Free Jupyter notebook environment
  - About 13G memory and 25G disk space limit
- No setup
- 100% Cloud
- Python + R
- Hardware accelerator
  - Graphical Processing Unit (GPU)
  - Tensor Processing Unit (TPU)

Awesome Sharing Feature

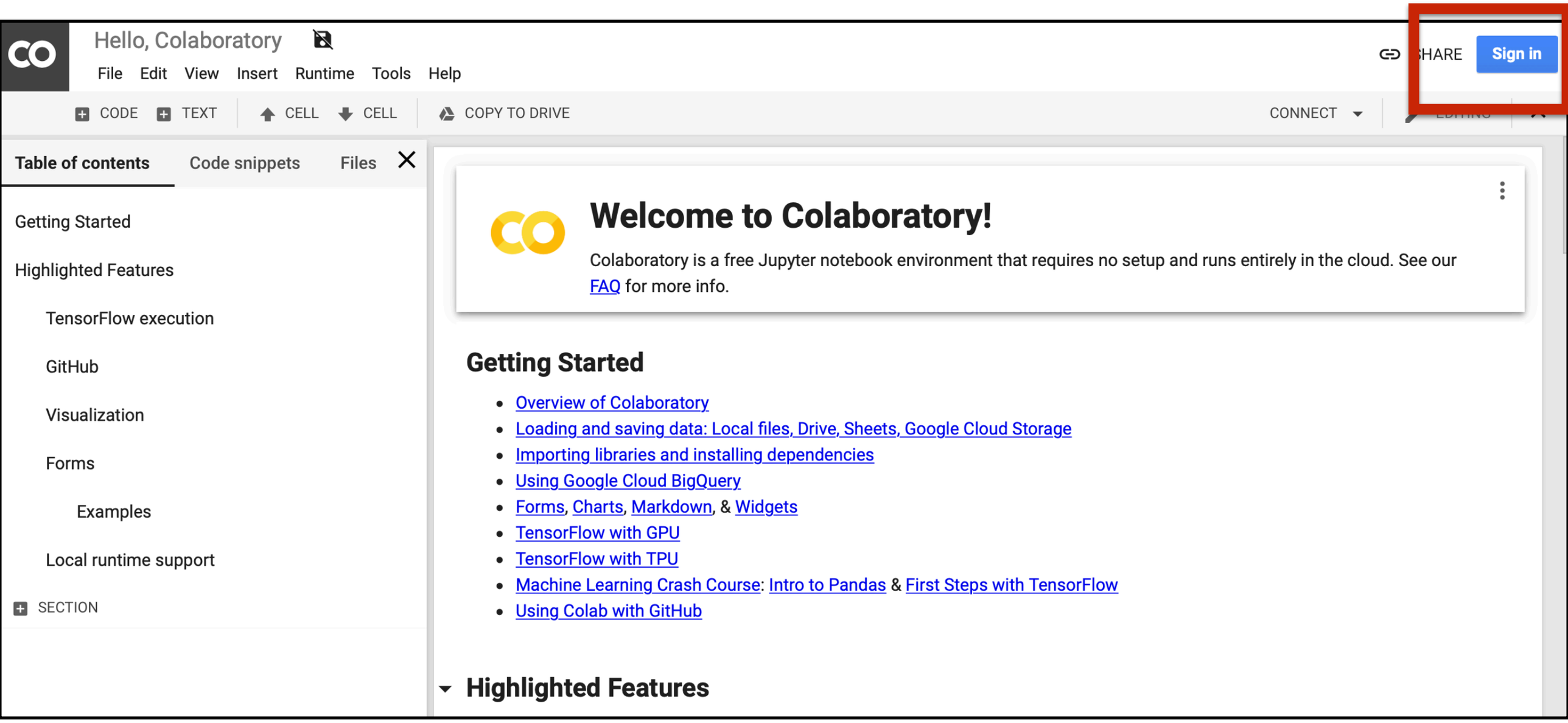
- Save Colab to Google Drive and share.
- Save Colab to GitHub directly.
- Upload / Download Colab to local drive.
- Share with others users
- Control revisions
- Google document
- Install packages

Cool Features

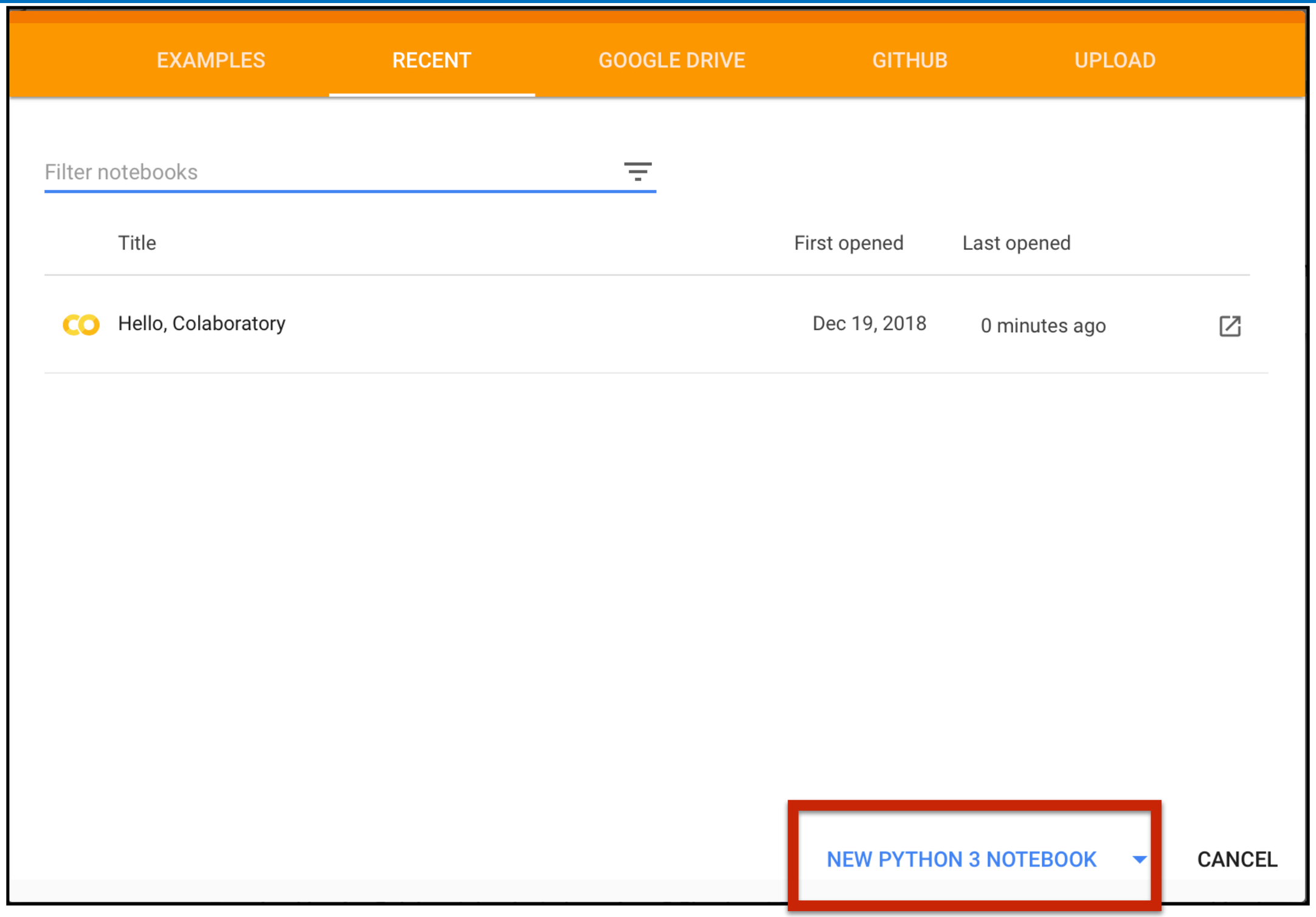
- If an error occurs, it (usually) automatically provides a link to StackOverflow
- You can add forms for input parameters from user.
- Mount Google Drive and access data from Google Sheet.

Google Colaboratory

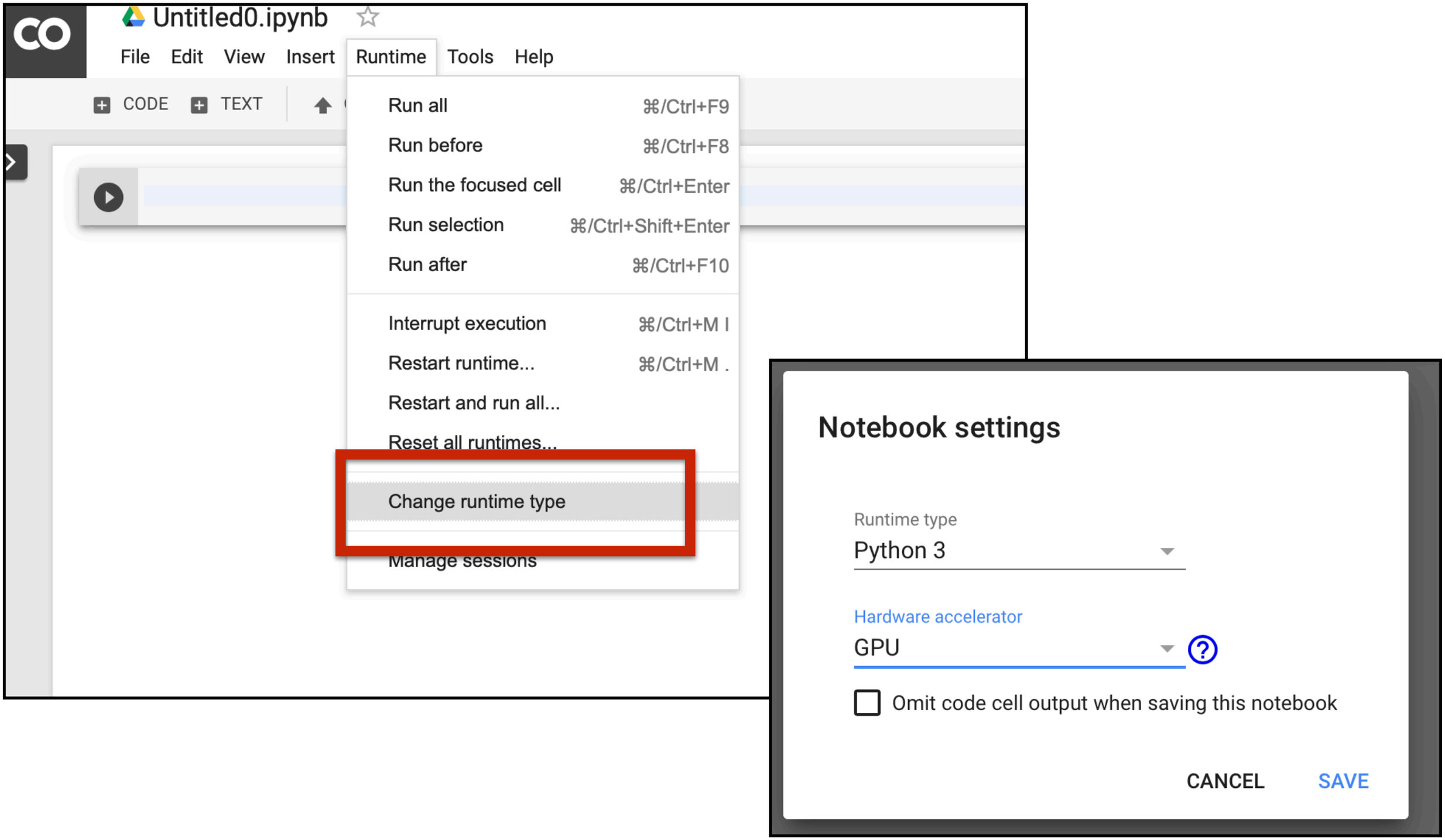
https://colab.research.google.com/notebooks/welcome.ipynb



Google Colaboratory



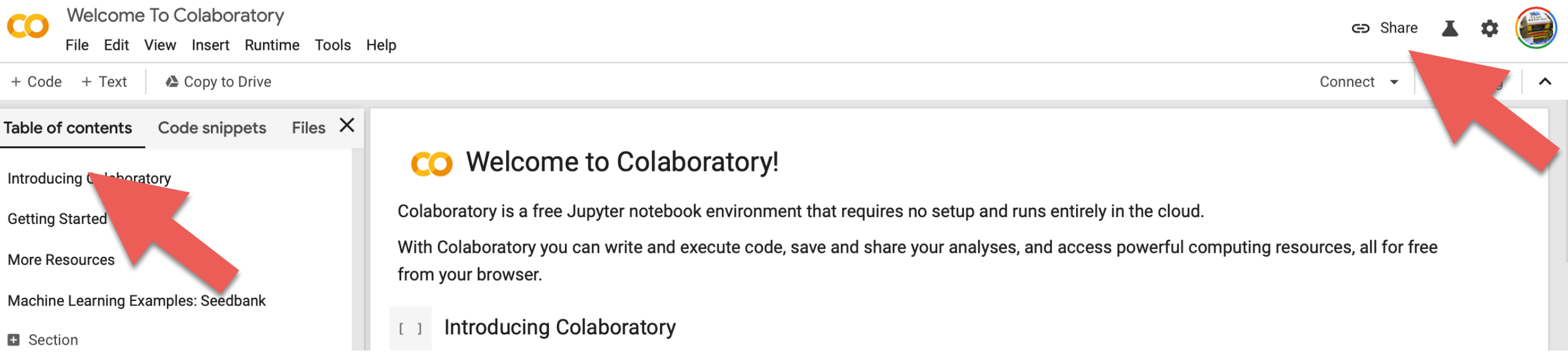
Google Colaboratory



Google Colaboratory R

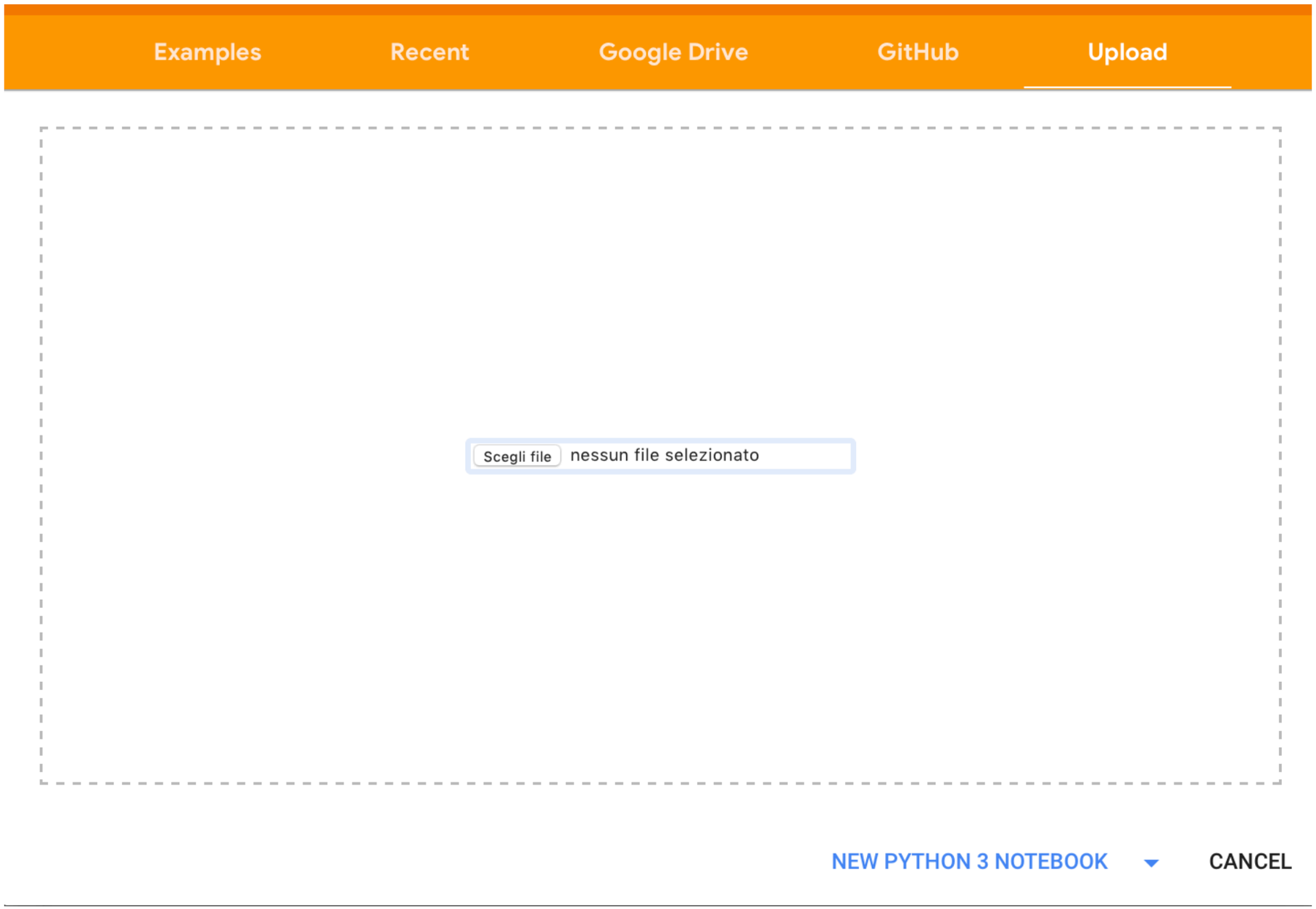
https://colab.research.google.com/notebook#create=true&language=r

Google Colaboratory R



Google Colaboratory R

Upload Demo_R notebook



Google Colaboratory R

Demo.ipynb

File Edit View Insert Runtime Tools Help

Comment Share

+ Code + Text

Initializing Editing

Code without visible output:

```
a <- 8
```

```
[ ] b <- 4:59
```

With visible output:

```
[ ] a + b
```

1. 12
2. 13
3. 14
4. 15

Databricks

Databricks

- E’ una piattaforma di analytics facile e collaborativa basata su **Apache Spark**

- Tutti i moduli di Spark sono presenti in Databricks (SparkSQL, Streaming, ML, GraphX)

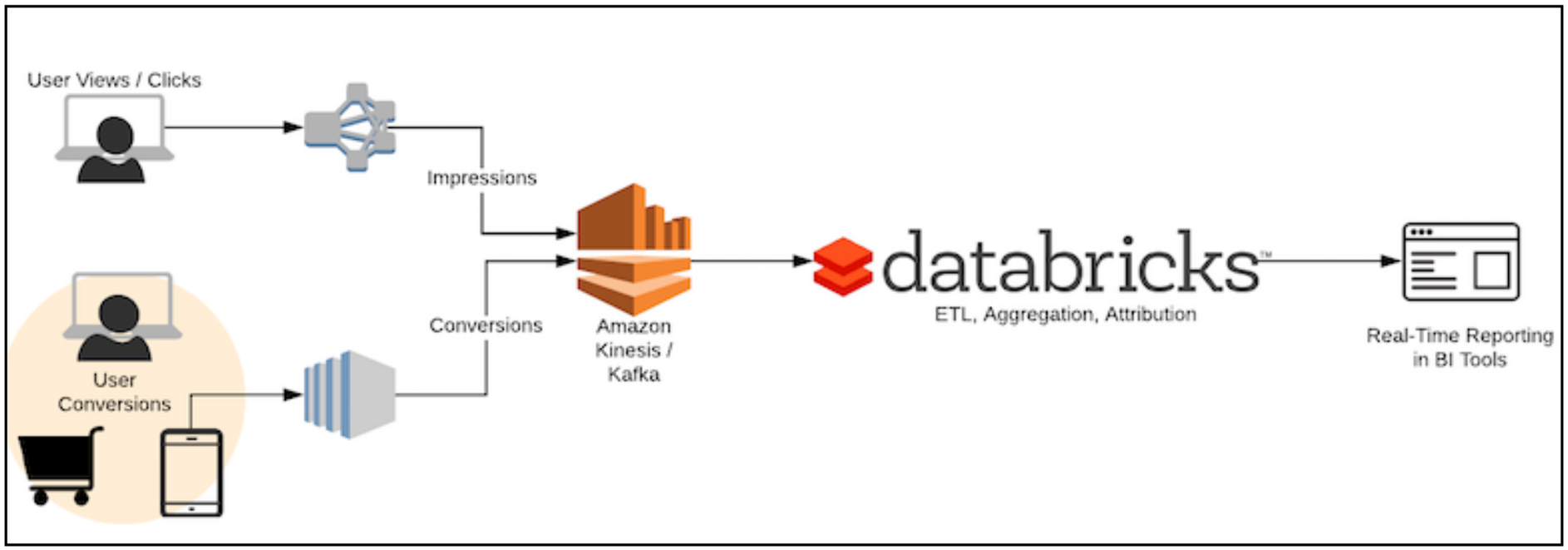
- **“Main goal: to remove all the hardness and complexity to get and manage a Spark cluster”**

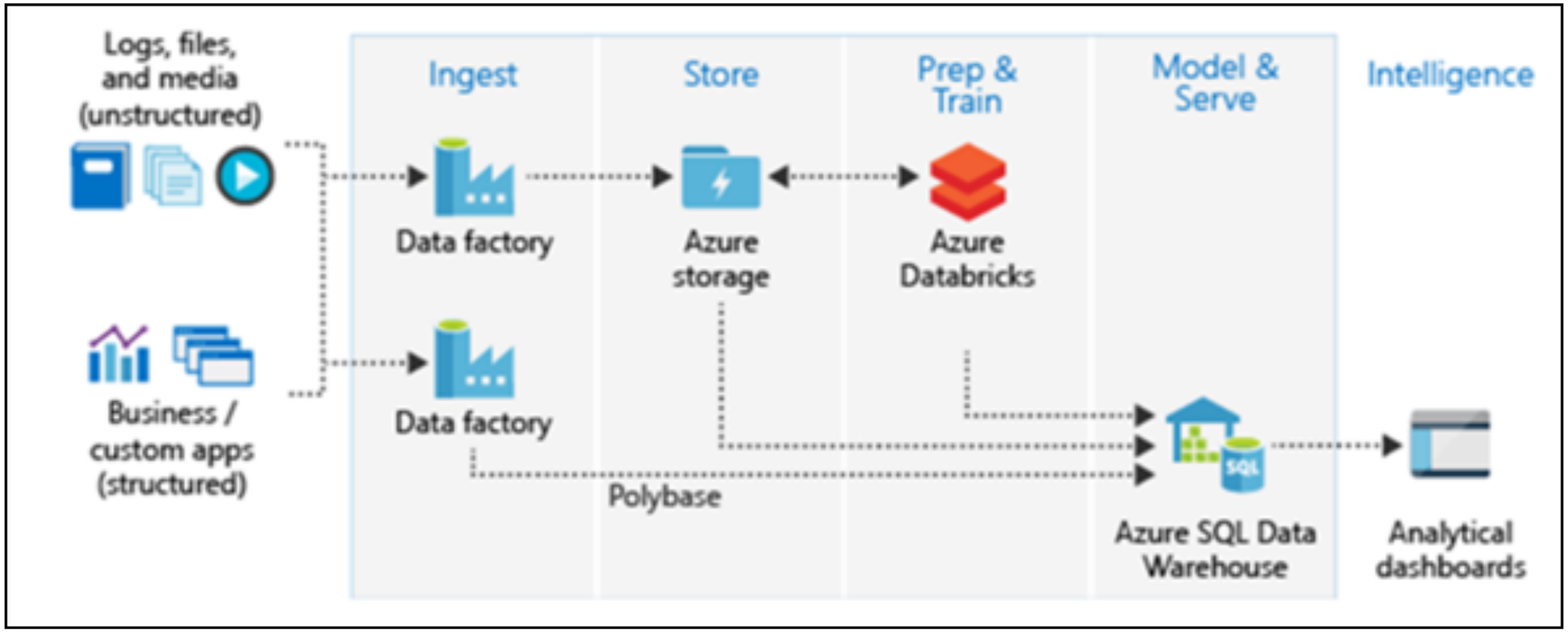
- Permette di gestire installazioni e impostazioni con un clic

- Offre flussi di lavoro semplificati e area di lavoro interattiva per facilitare la collaborazione tra data scientists, sviluppatori e business analysts

- Integrazione con i maggiori cloud providers come Amazon AWS and Microsoft Azure

Environments
AWS and Azure





Notebook

- Simile ai notebook di Jupyter o Zeppelin.Linguaggi supportati: R, Python, Scala e SQL

  - Possono essere usati tutti in un singolo notebook.

- La Spark session è gia definita per ciascun notebook come variabile globale spark.

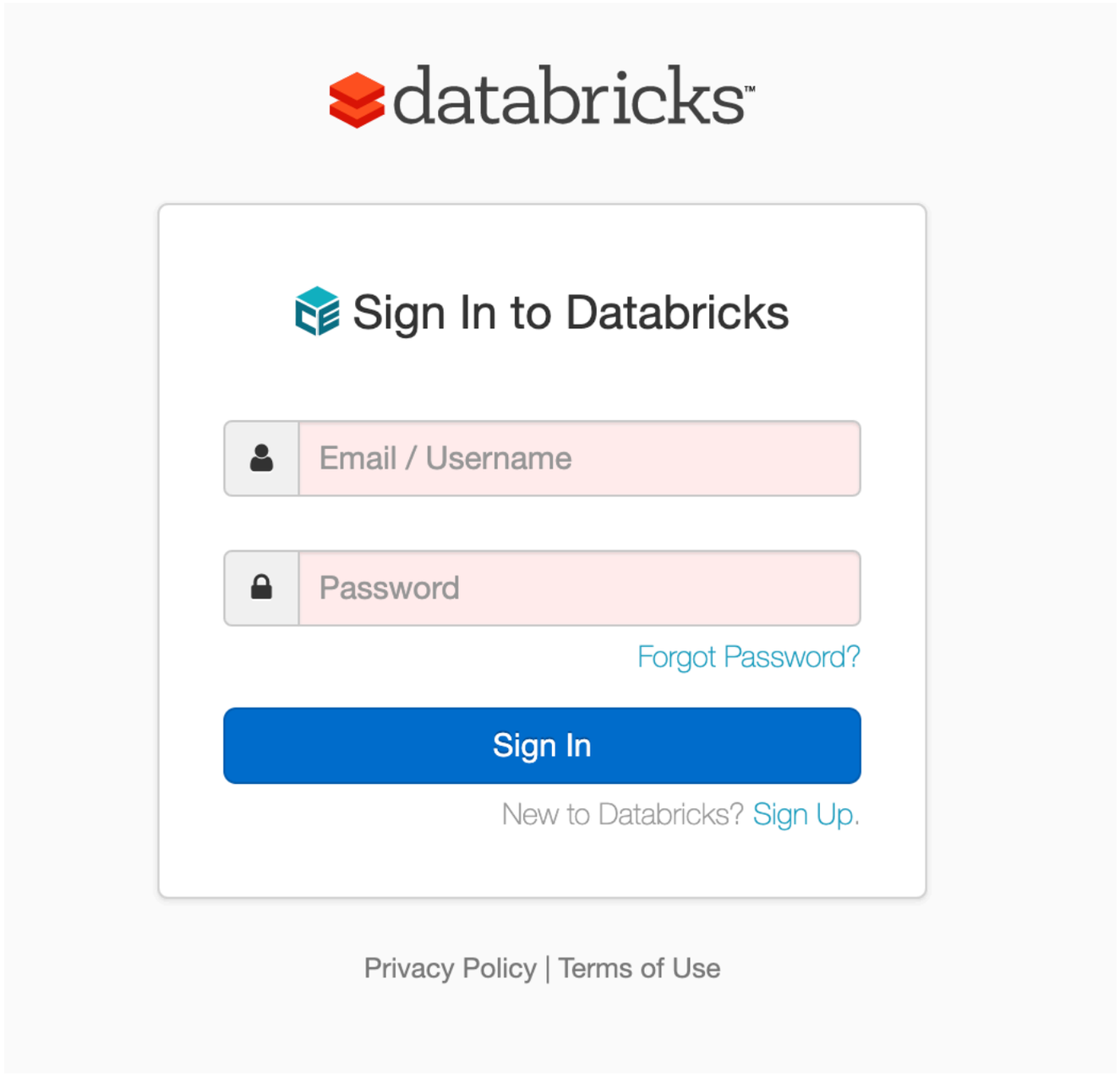
- Una volta creato un notebook deve essere collegato a un cluster attivo

Version Management

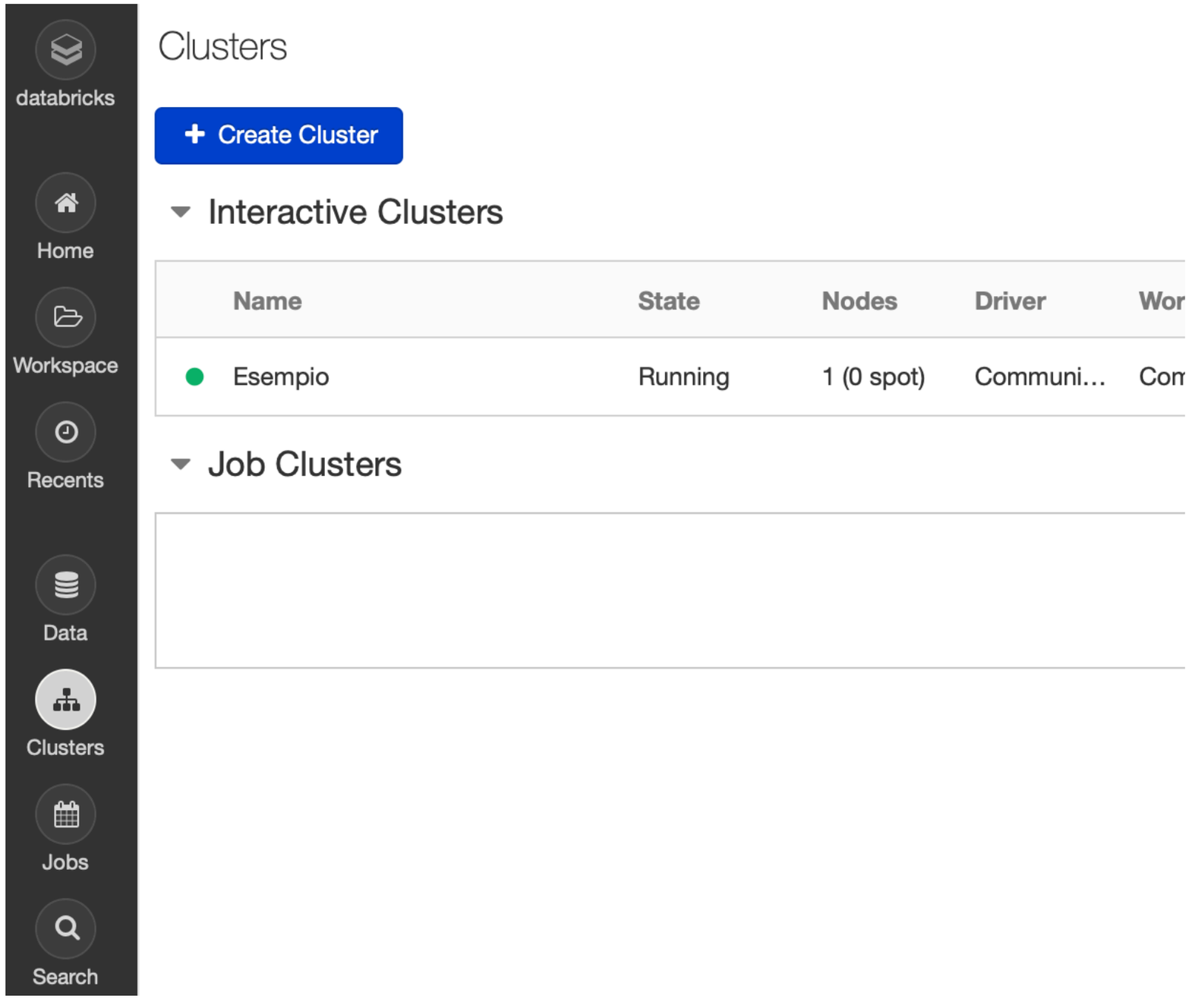
- Databricks é una piattaforma di analisi collaborativa dove gli utenti possono condividere workspaces, clusters and jobs attraverso una sola interfaccia.

- E’ possibile creare modelli condivisi nello stesso notebook real time, riutilizzare data assets, librerie presenti su uno stesso cluster, o riutilizzare/monitorare scheduled jobs.

- Databricks supporta l’integrazione con Github, Bitbucket Cloud & Azure DevOps Services



https://community.cloud.databricks.com/login.html


databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Clusters
+ Create Cluster
Interactive Clusters
Name
State
Nodes
Driver
Wor
Esempio
Running
1 (0 spot)
Communi…
Con
Job Clusters

Creazione del cluster

Create Cluster

New Cluster | Cancel | Create Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU

Cluster Name

Esempio1

Databricks Runtime Version

Runtime: 5.3 (Scala 2.11, Spark 2.4.1)

Python Version

3

Instance

Free 6GB Memory: As a C[illegible]munity Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.

Instances | Spark

Availability Zone

us-west-2c

us-west-2c

Upload dei dati

Create New Table

Upload File | S3 | DBFS | Other Data Sources

Upload to DBFS

/FileStore/tables/ (optional) Select

File

Team_Stats_PW_xlsx.csv

69.1 KB

Remove file

File uploaded to */FileStore/tables/Team_Stats_PW_xlsx-f88a9.csv*

Create Table with UI | Create Table in Notebook

Select a Cluster to Preview the Table

Choose a cluster with which you will read and preview the data.

Cluster

Esempio

Preview Table

Upload dei dati

Preview Table

Specify Table Attributes

Specify the Table Name, Database and Schema to add this to the data UI for other users to access

Table Name
team_stats_pw_xlsx_f88a9_

Create in Database
default

File Type
CSV

Column Delimiter
,

First row is header

Infer sc

Multi-line

Create Table

Create Table in Notebook

Table Preview

_c0	_c1	_c2	_c3	_c4
...G	STRING	STRING	STRING	STRING
GAME	LOCATION	COORD_X	COORD_Y	DISTANCE
vs Trieste R12	H	227.0	191.0	7.087384919700637
vs Trieste R12	H	180.0	60.0	7.995
Trieste R12	H	59.0	209.0	0.5418053155885424
vs Trieste R12	H	58.0	240.0	1.0296295450306387
vs Trieste R12	H	89.0	214.0	1.639856091247034
vs Trieste R12	H	147.0	236.0	3.9772352206023722

Upload dei dati

Create New Table

Preview Table

Specify Table Attributes

Specify the Table Name, Database and Schema to add this to the data UI for other users to access

Table Name

team_stats_pw_xlsx_f88a9_

Create in Database

default

File Type

CSV

Column Delimiter

,

First row is header

Infer s

Multi-line

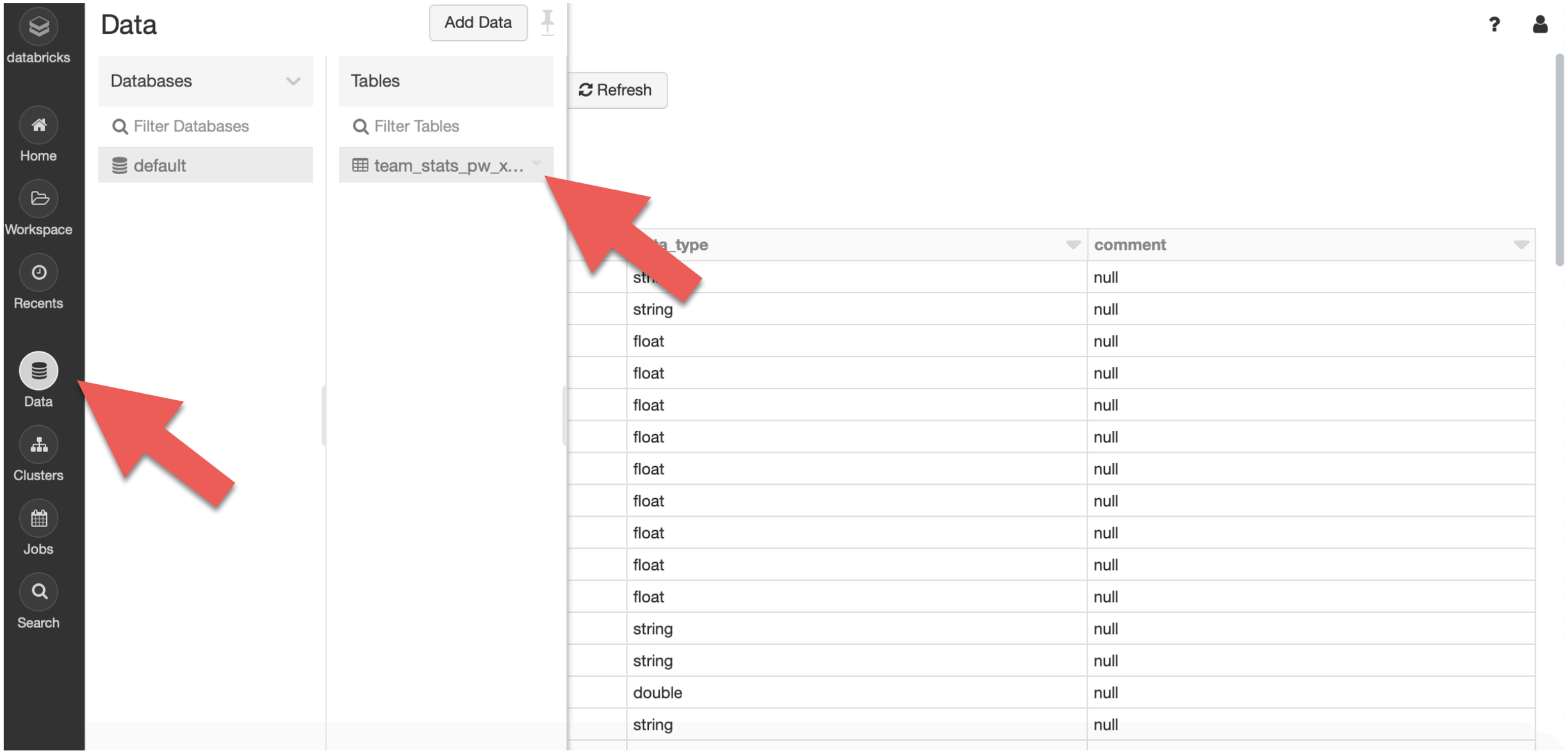
Create Table

Create Table in Notebook

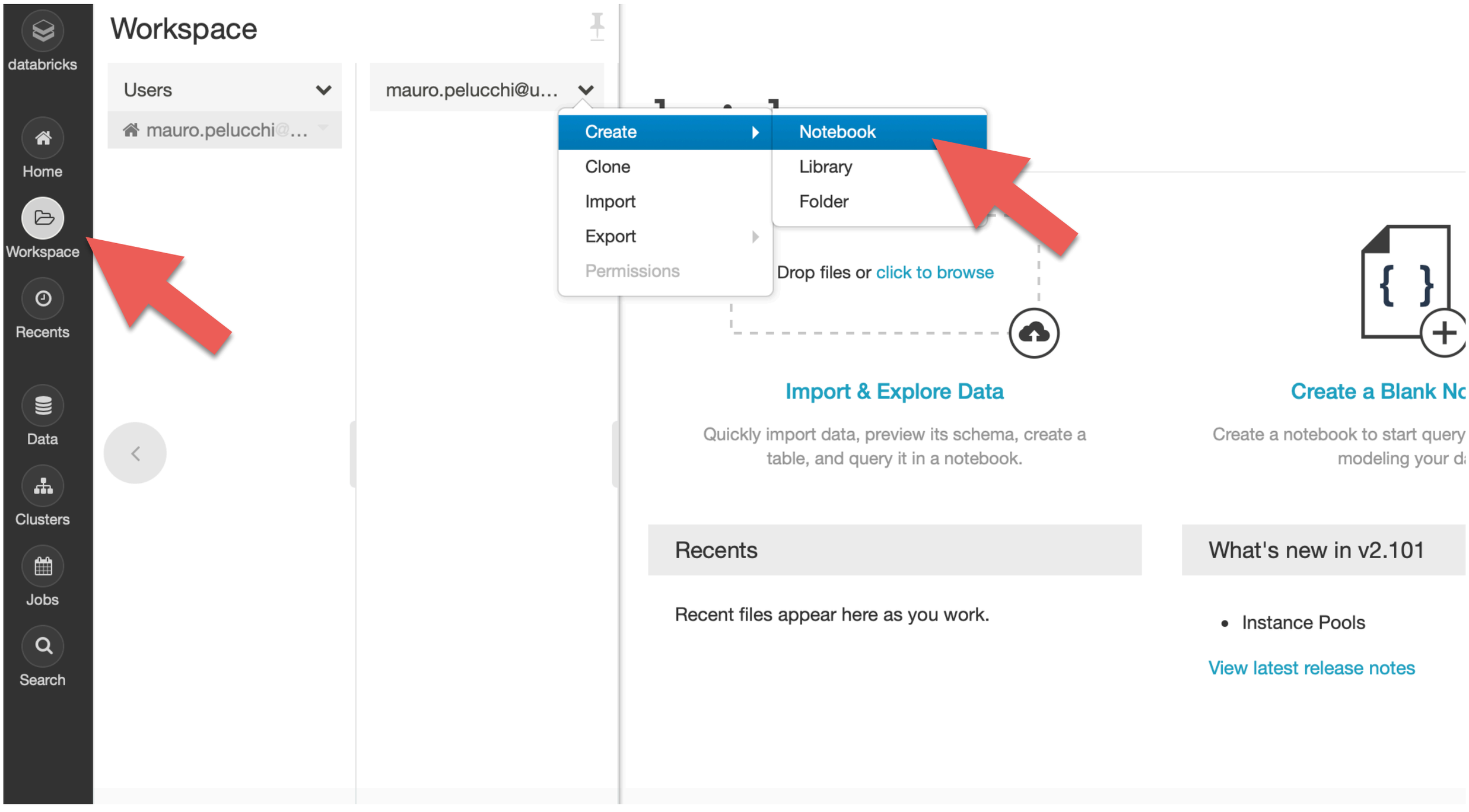
Table Preview

GAME	LOCATION	COORD_X	COORD_Y	DISTANCE
STRING	STRING	FLOAT	FLOAT	FLOAT
vs Trieste R12	H	227.0	1.0	7.087384919700637
vs Trieste R12	H	180.0	60.0	7.995
vs Trieste R12	H	59.0	209.0	0.5418053155885424
vs Trieste R12	H	58.0	240.0	1.0296295450306387
vs Trieste R12	H	89.0	214.0	1.639856091247034
vs Trieste R12	H	147.0	236.0	3.9772352206023722
vs Trieste R12	H	236.0	243.0	7.445834405894345

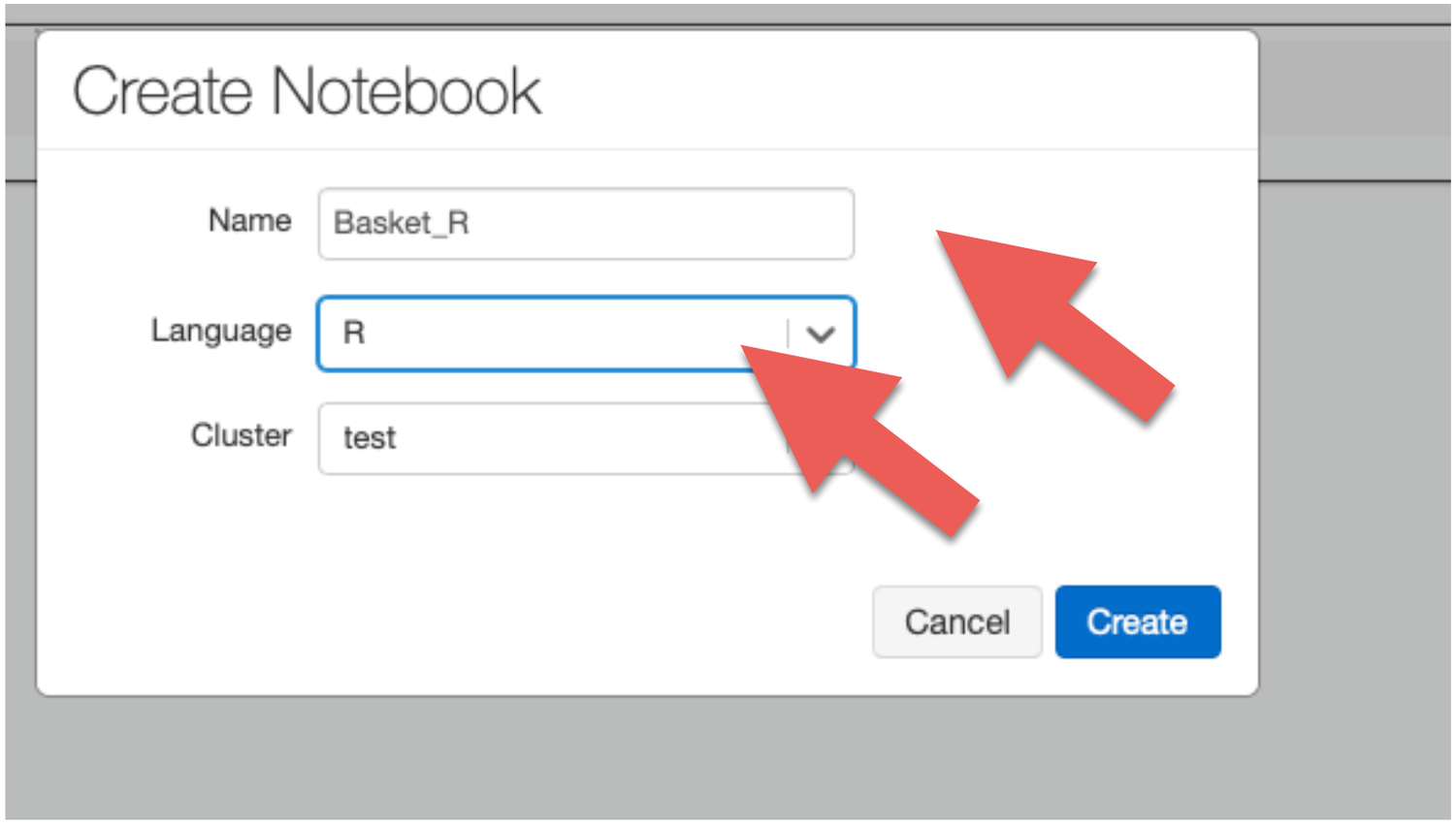
Database


databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Data
Add Data
Databases
Filter Databases
default
Tables
Filter Tables
team_stats_pw_x...
Refresh
comment
null
string
null
float
null
float
null
float
null
float
null
float
null
float
null
float
null
string
null
string
null
double
null
string
null

Notebook


databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Workspace
Users
mauro.pelucchi@u…
mauro.pelucchi@…
Create
Clone
Import
Export
Permissions
Notebook
Library
Folder
Drop files or click to browse
Import & Explore Data
Quickly import data, preview its schema, create a table, and query it in a notebook.
Create a Blank No
Create a notebook to start query
modeling your da
Recents
Recent files appear here as you work.
What's new in v2.101
Instance Pools
View latest release notes

Notebook


Create Notebook
Name
Basket_R
Language
R
Cluster
test
Cancel
Create

Notebook

https://docs.databricks.com/notebooks/notebooks-use.html#develop-notebooks

Notebook - R

Basket_R (R)

test | File | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

Cmd 1

```
1
2
3 library(dplyr)
4 library(caret)
5 library(data.table)
```

```
Loading required package: lattice
Loading required package: ggplot2

Attaching package: 'data.table'

The following objects are masked from 'package:dplyr':

    between, first, last
```

Command took 2.51 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 16:29:06 on test

Shift+Enter to run shortcuts

Notebook —> SQL

Cmd 2

```
1  %sql
2
3  select * from defau  m_stats_1
```

▸ (1) Spark Jobs

GAME	LOCATION	COOR A	D_Y	DISTANCE	SHOT_CLOCK	DRIBBLES	TOUCHES	AREA	ASSISTS	EXTRA_PASS	FROM	END	TARGET	SITUATION	FIN/
vs Trieste R12	H	227	191	7.087384919700637	11	12	4	0	0	0	SHOT_IN	3_SHOT_MISSED	0	P	W
vs Trieste R12	H	180	60	7.995	6	18	5	0	0	0	FT_IN	3_SHOT_MISSED	0	CO	W
vs Trieste R12	H	59	209	0.5418053155885424	21	3	1	1	1	0	STEAL	2_SHOT_MADE	1	FB	W
vs Trieste R12	H	58	240	1.0296295450306387	8	16	4	1	0	0	STEAL+THROW-IN	2_SHOT_MISSED	0	P+I	W
vs Trieste R12	H	89	214	1.639856091247034	23	2	1	1	0	0	STEAL	X_FS	0	FB	W

Command took 1.38 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 16:30:20 on test

SparkR

Cmd 3

```
1 library(SparkR)
2 basket_data <- sql("select * from default.team_stats_1")
3 display(basket_data)
4
```

▸ (1) Spark Jobs

BLES	TOUCHES	AREA	ASSISTS	EXTRA_PASS	FROM	END	TARGET	SITUATION	FINAL	PERIOD	SCORE	DIFFERENCE	PLAYER	Column	LINK	CHAMP_LE
	4	0	0	0	SHOT_IN	3_SHOT_MISSED	0	P	W	1	0--2	-2	Phil Greene IV	null	team/vs Trieste R12/1	A2
	5	0	0	0	FT_IN	3_SHOT_MISSED	0	CO	W	1	0--4	-4	Jamal Jones	null	team/vs Trieste R12/2	A2
	1	1	1	0	STEAL	2_SHOT_MADE	1	FB	W	1	0--4	-4	Jamal Jones	null	team/vs Trieste R12/3	A2
	4	1	0	0	STEAL+THROW-IN	2_SHOT_MISSED	0	P+I	W	1	2--4	-2	Curtis Nwohuocha	null	team/vs Trieste R12/4	A2
	1	1	0	0	STEAL	X_FS	0	FB	W	1	2--4	-2	Phil Greene IV	null	team/vs Trieste R12/5	A2

Command took 1.33 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 17:46:50 on test

Databricks

Cmd 4

```
counts <- count(groupBy(basket_data, basket_data$PLAYER))
display(counts)
```

▸ (5) Spark Jobs

PLAYER	count
Mitchell Poletti	20
Riccardo Visconti	15
Simone Pierich	15
Andrea Amato	79
Mattia Udom	60
Leonardo Tote	19
Omar Dieng	1
Curtis Nwohuocha	9
Phil Greene IV	87

Command took 1.83 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 17:48:24 on test

Shift+Enter to run shortcuts

Databricks

Cmd 4

```
1 counts <- count(groupBy(basket_data, basket_data$PLAYER))
2 display(counts)
```

▸ (5) Spark Jobs

PLAYER	count
Mitchell Poletti	20
Riccardo Visconti	15
Simone Pierich	15
Andrea Amato	79
Mattia Udom	60
Leonardo Tote	19
Omar Dieng	1
Curtis Nwohuocha	9
Phil Greene IV	

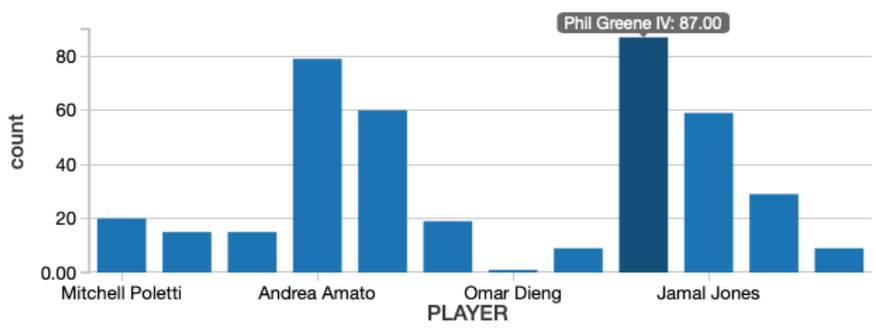
Command took 1.83 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 17:48:24 on test

Shift+Enter to run shortcuts

Cmd 4

```
1 counts <- count(groupBy(basket_data, basket_data$PLAYER))
2 display(counts)
```

▸ (5) Spark Jobs



Plot Options...

Command took 1.83 seconds -- by mauro.pelucchi@unimib.it at 17/1/2020, 17:48:24 on test

Shift+Enter to run shortcuts