

**IBM Think 2018**

**Hands-on**

**Qualita’ e Controllo Costi:**

**un connubio possibile**

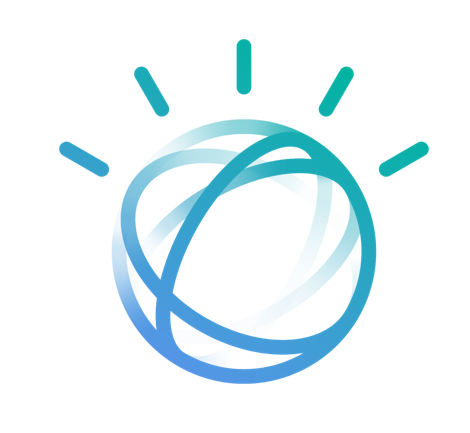
## Daniela Zuppini, IBM, [daniela\_zuppini@it.ibm.com](mailto:daniela_zuppini@it.ibm.com)

Ludovica Scarfi’, IBM,

## [ludovica.scarfi@it.ibm.com](mailto:ludovica.scarfi@it.ibm.com)

## Margherita de Cesare, IBM, [margherita.decesare@it.ibm.com](mailto:margherita.decesare@it.ibm.com)

## Dino Ravasi, IBM, [dino\_ravasi@it.ibm.com](mailto:dino_ravasi@it.ibm.com)



**Indice**

[Introduzione 3](#_Toc515548946)

[Prerequisiti 5](#_Toc515548947)

[Ingredienti 5](#_Toc515548948)

[Registrazione a IBM Cloud 5](#_Toc515548949)

[Set up di Watson Studio 6](#_Toc515548950)

[Set up di un progetto in Watson Studio 8](#_Toc515548951)

[Set up di Apache Spark in Watson Studio 11](#_Toc515548952)

[Lavoriamo con i Notebook 12](#_Toc515548953)

[Creazione di un Notebook in Watson Studio 12](#_Toc515548954)

[Import del dataset nel Notebook in Watson Studio 15](#_Toc515548955)

[Esecuzione nel Notebook in Watson Studio 17](#_Toc515548956)

[Reference 18](#_Toc515548969)

# **Introduzione**

La Sanita' e' un ambito ricco di dati ma questi dati sono complessi da gestire e per varie ragioni:

* normative per la tutela della privacy
* ampia dimensionalita' dei dati: si parla di migliaia di patologie, farmaci e protocolli di cura
* presenza di dati non strutturati: ad esempio le annotazioni del personale medico
* mancanza di standard e molteplicita' di piattaforme: strutture sanitarie diverse useranno standard e tecnologie diverse.

Ripercorriamo quanto realizzato da D2hawkeye, una societa' americana specializzata in analisi dati del settore medico, con la collaborazione del MIT, Massachusetts Institute of Technology. Siamo quindi nell'ambito del sistema sanitario USA: il paziente si rivolge ad una struttura sanitaria per effettuare visite mediche, sottoporsi a esami e terapie, ottenere la prescrizione di farmaci; in una fase successiva, la struttura sanitaria si rivolgera' alla compagnia assicurativa del paziente per ottenere il rimborso dei costi sostenuti.

Esaminando i dati relativi alle richieste rimborso costi, la D2hawkeye ha messo a punto un sistema in grado di valutare la qualita’ dei percorsi di cura corrispondenti.

Poter effettuare uno screening dei percorsi di cura in modo automatico e su larga scala, consente di:

* evidenziare pazienti non trattati correttamente e intraprendere azioni correttive
* ottenere un migliore utilizzo delle risorse del sistema sanitario con conseguente risparmio dei costi.

**Obiettivi del Laboratorio**

In questo laboratorio costruiremo un modello che, esaminando i dati delle richieste rimborso costi sanitari, e’ in grado di valutare se il percorso di cura proposto al paziente e’ stato quello piu’ appropriato, o al contrario non risponde in modo ottimale alla patologia.

Scriveremo un programma il cui codice non avra' lo scopo di definire in forma di source code le regole in base alle quali rispondere a degli input. Al contrario il codice che scriveremo avra' lo scopo di imparare dai dati, in quanto rappresentazione dell’esperienza passata, le regole in base alle quali rispondere agli input futuri.

Il metodo che utilizzeremo si chiama Regressione Logistica e la useremo per effettuare una classificazione binomiale, cioe’ suddivideremo i dati in due classi distinte.

In particolare andiamo a separare i pazienti che sono stati trattati in modo non appropriato dal resto dei pazienti. Per fare questo useremo una variabile, la chiameremo PoorCare, che assumera’ i valori:

* PoorCare=TRUE per indicare una scarsa qualita’ del percorso di cura
* PoorCare=FALSE per indicare una buona qualita’ del percorso di cura

Il modello sara’ in grado di associare ad ogni record del data set la probabilita’ di appartenere a ciascuna delle due classi.

Per mettere a punto il modello scriveremo un Jupyter Notebook in linguaggio R.

Il nostro codice sara' focalizzato sugli aspetti di lettura dati, sulla comprensione di quali parametri sono rilevanti per il problema, nell’analizzare e decidere il livello di soglia da applicare alla probabilita’ in modo da attribuire un valore TRUE/FALSE alla variabile PoorCare.

I servizi usati in questo tutorial sono: Apache Spark, Jupyter notebooks in linguaggio R all’interno della piattaforma IBM Watson Studio e IBM Cloud Object Storage per memorizzare dati e contenuto dei notebook.

Il dataset utilizzato in questo laboratorio e’ stato messo a punto da MIT. Di seguito il link per leggere di D2Hawkeye, MIT e del Prof. Bertsimas:

http://news.mit.edu/2016/startup-optimizes-corporate-health-care-plans-0218

**Prerequisiti**

**Skill Level:** Basico

**Piattaforma:** PC con un browser e una connessione internet

**Ingredienti**

1. **Accesso a IBM Cloud**: nel catalogo di IBM Cloud sono a disposizione servizi IaaS e PaaS. Useremo la componente PaaS e faremo attivazione di servizi in modalita’ lite a costo zero.
2. **IBM Watson Studio**: questo servizio e’ una piattaforma collaborativa per la preparazione di modelli analitici avanzati. E’ possibile scrivere modelli usando Scala, Python, R ed e’ possibile usare framework come Scikit-learn, Tensorflow e Keras.La piattaforma consente anche di utilizzare un Modeler Flows per scrivere modelli appoggiandosi a Watson Machine Learning Modeler e per costruire Neural Network usando un Neural Network Modeler.
3. **IBM Apache Spark**:questo servizio fornisce un ambiente Apache Spark semplicemente con un click. Con Spark e’ possibile eseguire operazioni sui dati in modalita’ parallela. Faremo attivazionedi un servizio Apache Spark specificando il piano lite che prevede due Spark executor.
4. **IBM COS**: Cloud Object Storage e’ un object storage basato sullo standard S3 che e’ viene usato da Watson Studio per salvare dati e notebook. Nel piano lite sono compresi 25GB di IBM COS a costo zero.

**Registrazione a IBM Cloud**

Per registrarti ad IBM Cloud puoi utilizzare il promo code della Cognitive Class:

[https://cognitiveclass.ai/ibm-cloud-promotion/](https://cognitiveclass.ai/ibm-cloud-promotion/%20)

Usando questo link puoi iscriverti ad un portale di corsi on-line gratuiti e specializzati su Data Science, Big Data, Machine Learning, Deep Learning. Insieme ai corsi viene fornito un promo code che fornisce un credito di 1200$ di risorse IBM Cloud.

[Clicca qui](http://cocl.us/CC_SIGNUP_CLOFF) (<https://courses.cognitiveclass.ai/login>) per registrarti e fare login alla Cognitive Class.

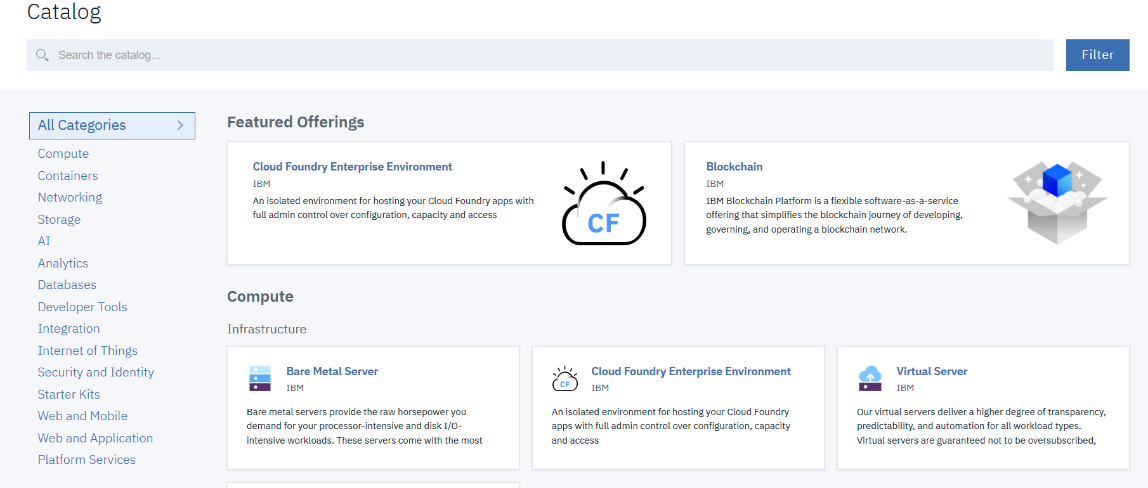
Ottieni il tuo promo code nella Welcome page (coloro che sono gia’ iscritti alla CognitiveClass trovano il promo code nella loro [Dashboard](https://courses.cognitiveclass.ai/dashboard))

[Registrati ad IBM Cloud](http://cocl.us/IBM_CLOUD_PROMO) e applica il tuo promo code

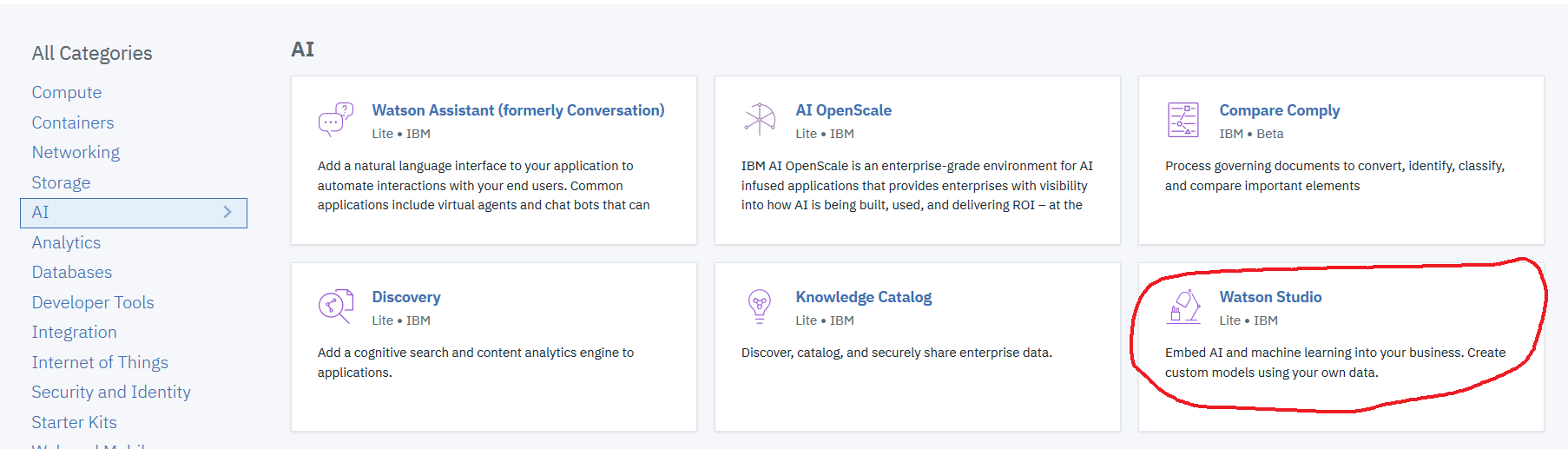
Se sei gia registrato fai Login.

**Set up di Watson Studio**

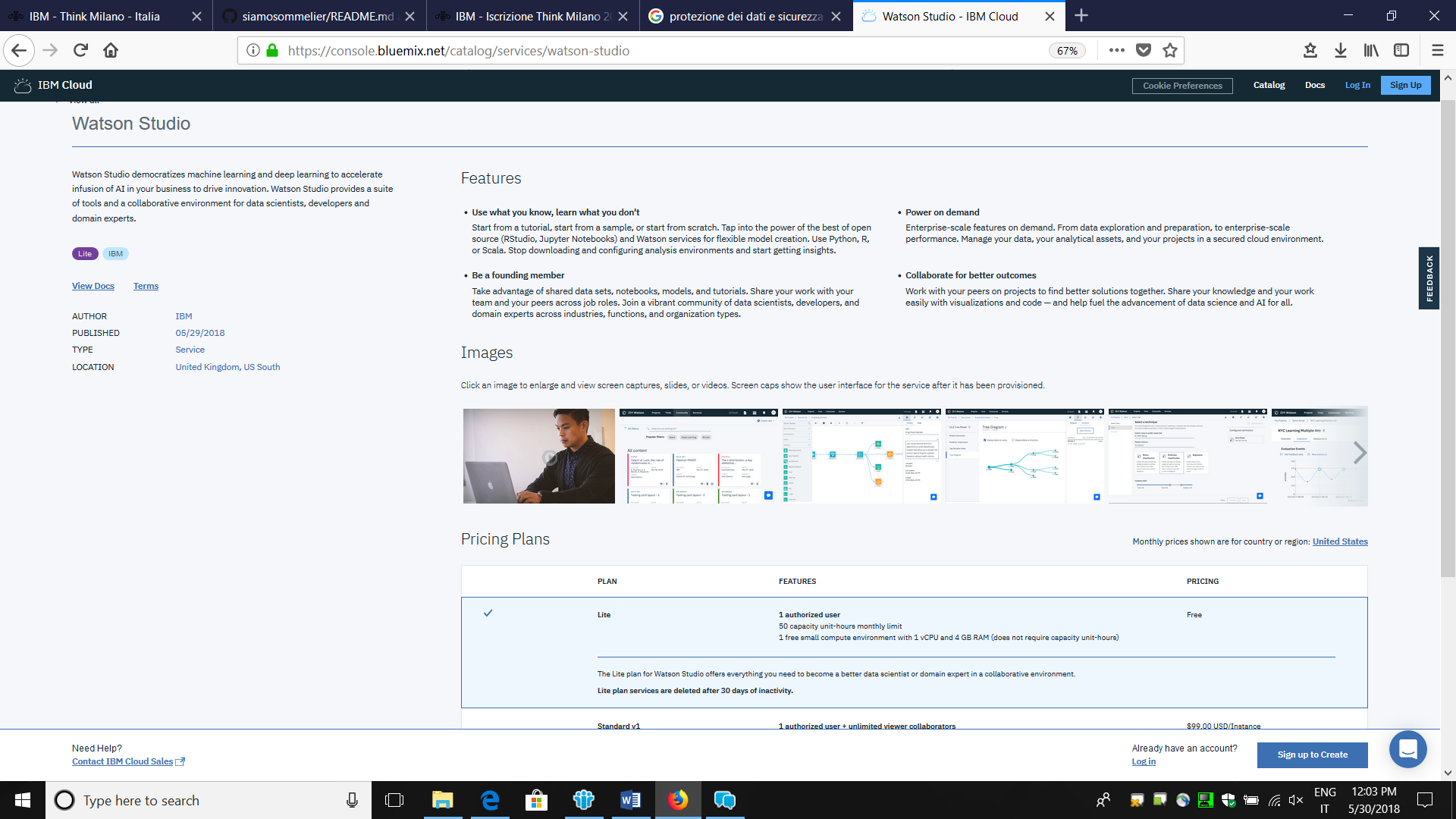
Dalla home page di IBM Cloud seleziona il Catalogo.



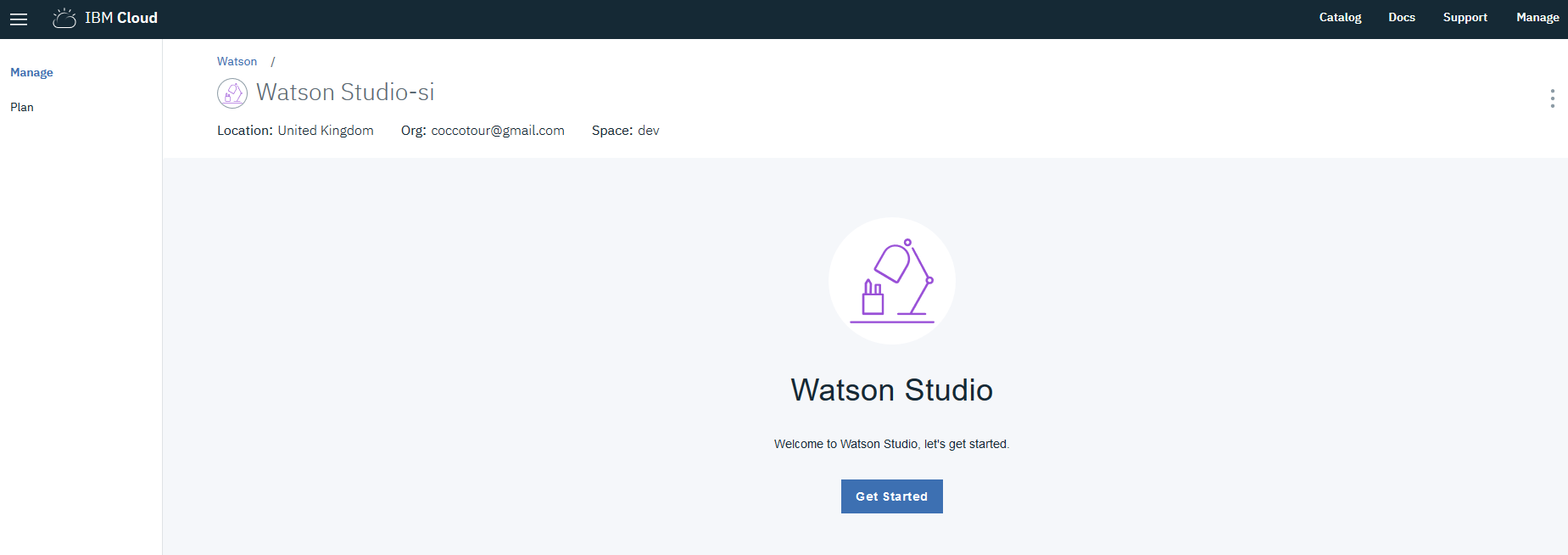
1. Nella sezionea sinistra dello schermo, seleziona AI.
2. Scegli il servizio Watson Studio facendo doppio click.



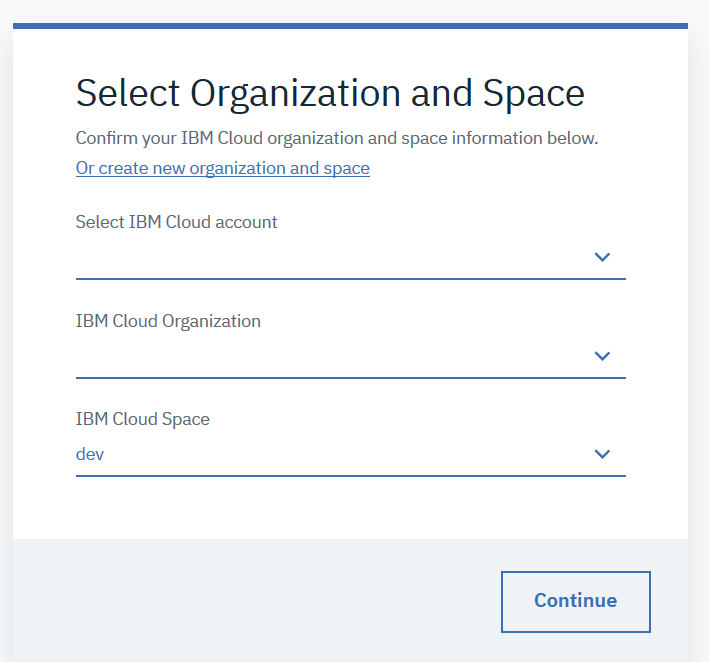
Quando si apre la pagina del servizio Watson Studio verifica che sia selezionato il piano lite e seleziona Create:



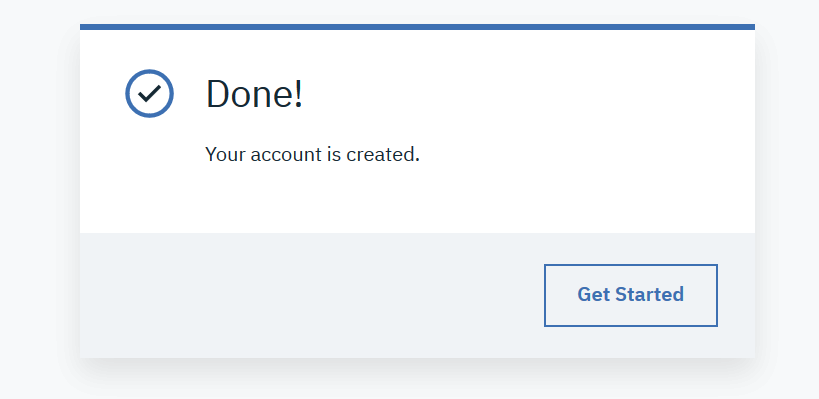
Quando il servizio Watson e’ partito, nella pagina principale di Watson Studio clicca sul tasto **GET STARTED**



Al primo accesso viene chiesto di confermare o modificare le informazioni di accesso allo spazio cloud. Scegli di confermare i valori proposti.

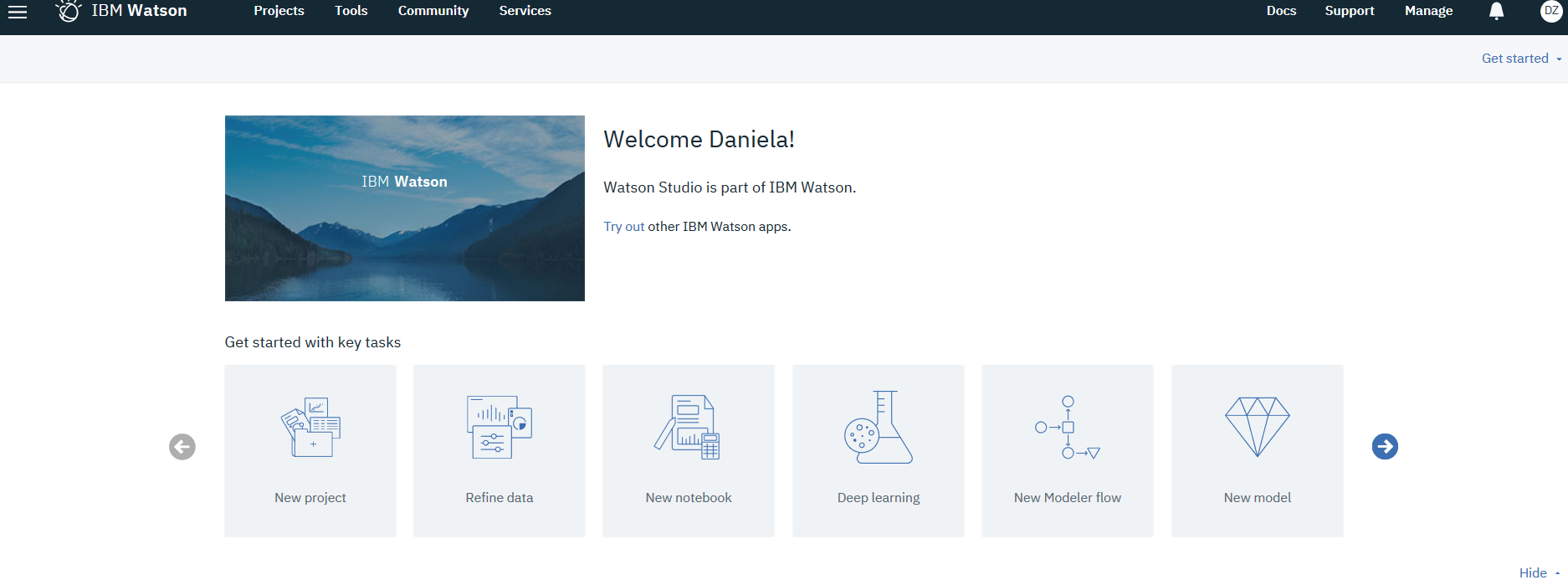


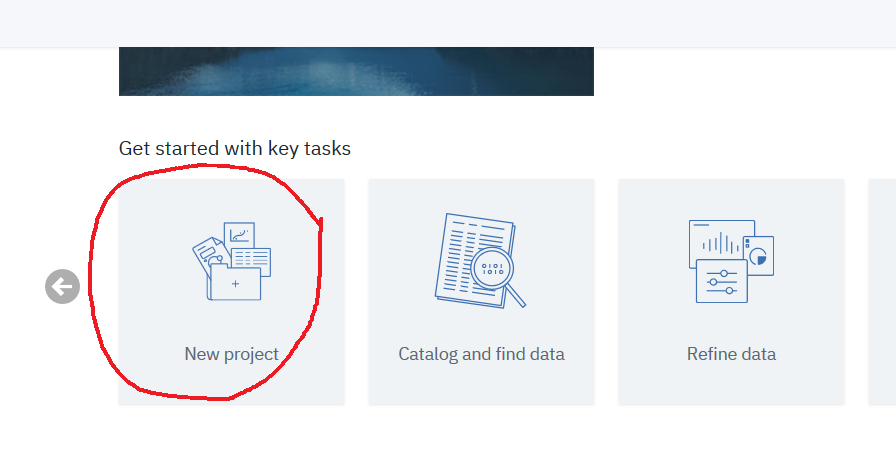
Il tuo account viene correttamente creato:



**Set up di un progetto in Watson Studio**

Nella pagina principale di Watson Studio:



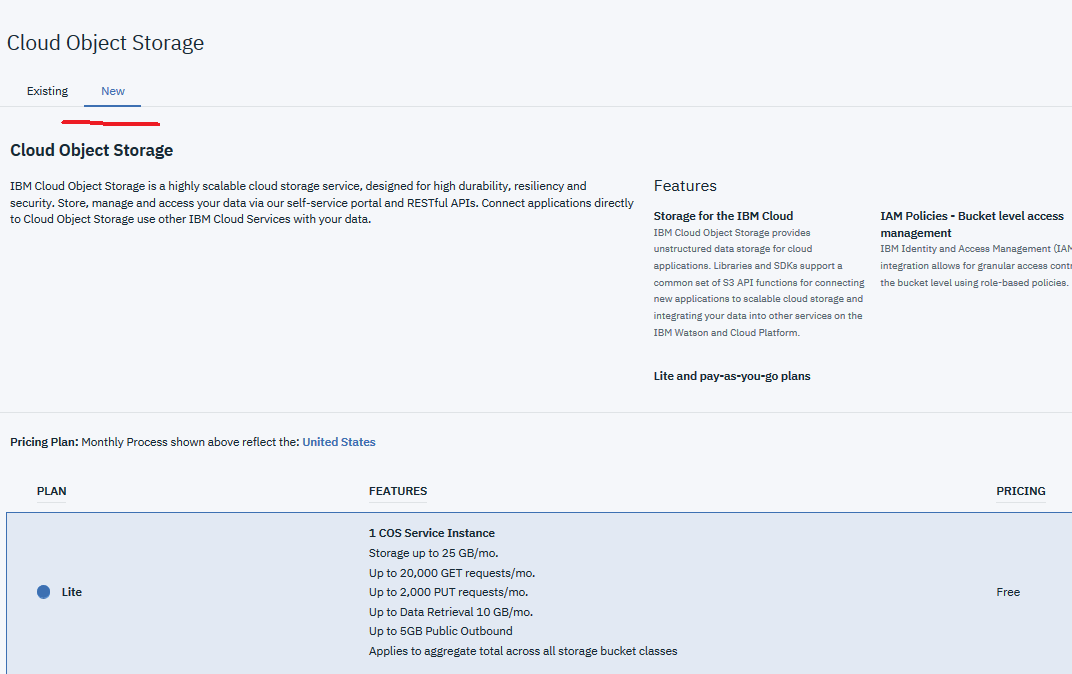
Crea un nuovo progetto e accetta la proposta di attivare tutti i servizi.

Se hai appena creato il servizio di Watson Studio, e questo e’ il tuo primo progetto, devi attivare anche un servizio di IBM Cloud Object Storage (COS).

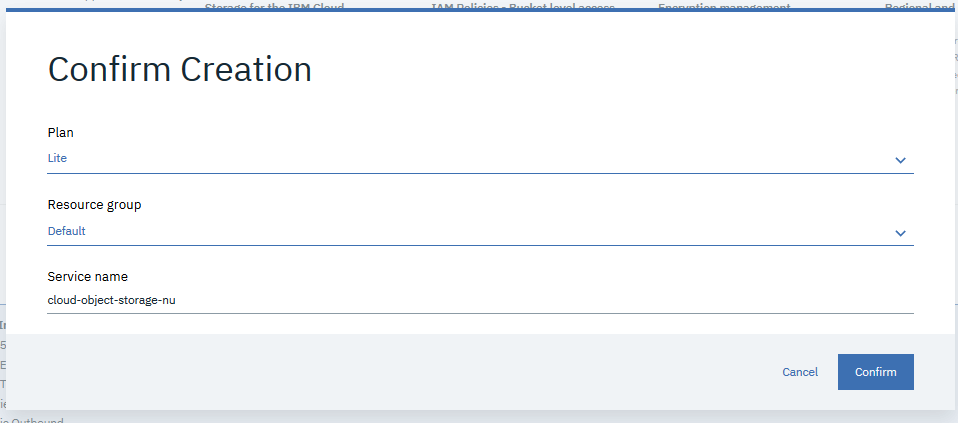
Scegli Add per attivare il servizio di IBM COS.



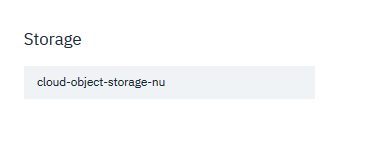
A questo punto si apre il servizio di Cloud Object Storage. Verifica che sia stato selezionato il piano lite e clicca su **create** per creare il servizio.



Conferma la creazione del servizio:



Fai click su refresh del pannello creazione del progetto. Comparira’ il nome dell’object storage appena creato nel cloud:

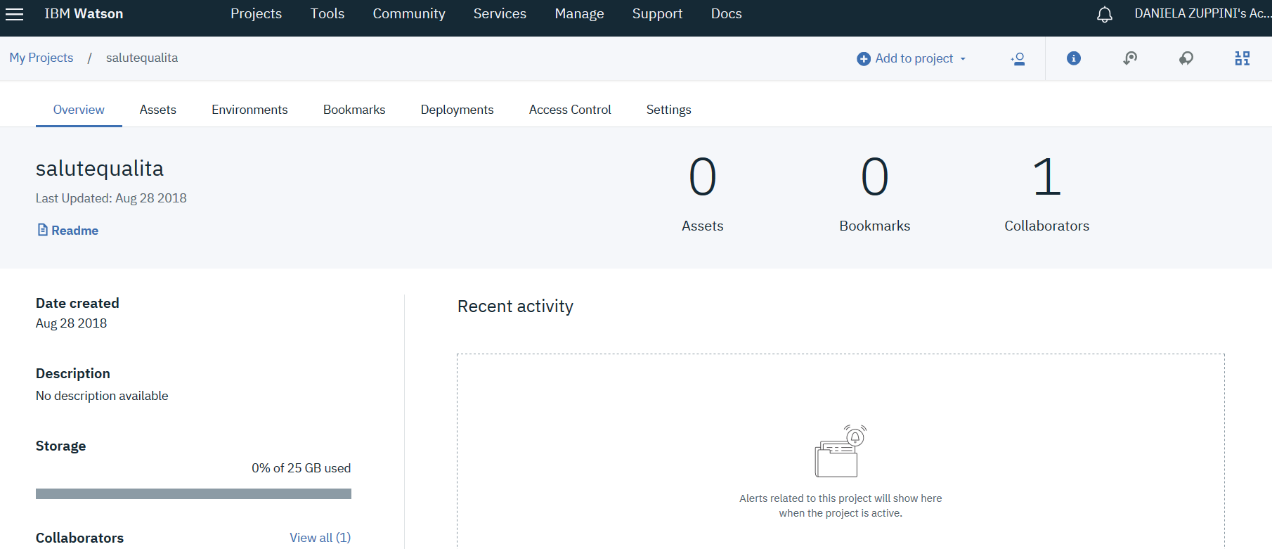


Diamo un nome al nuovo progetto e scegliamo **create**:

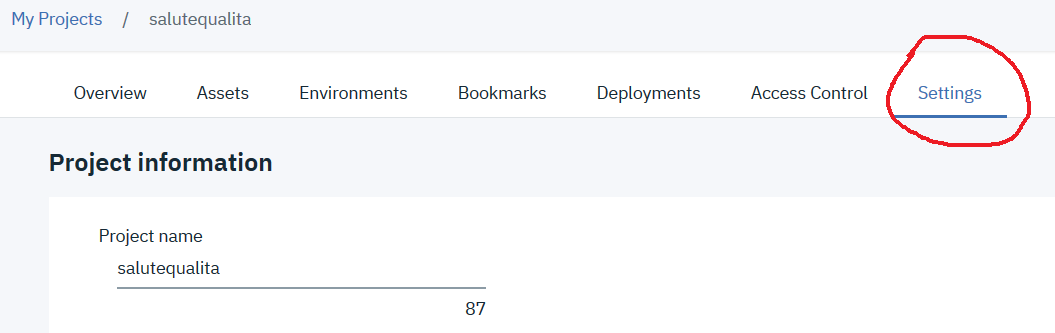


**Set up di Apache Spark in Watson Studio**

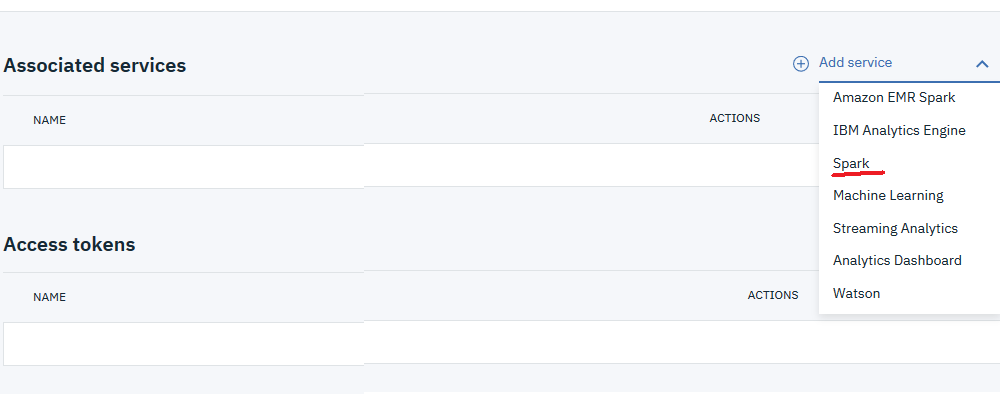
Quando il progetto viene creato, si apre la finestra di lavoro del progetto:



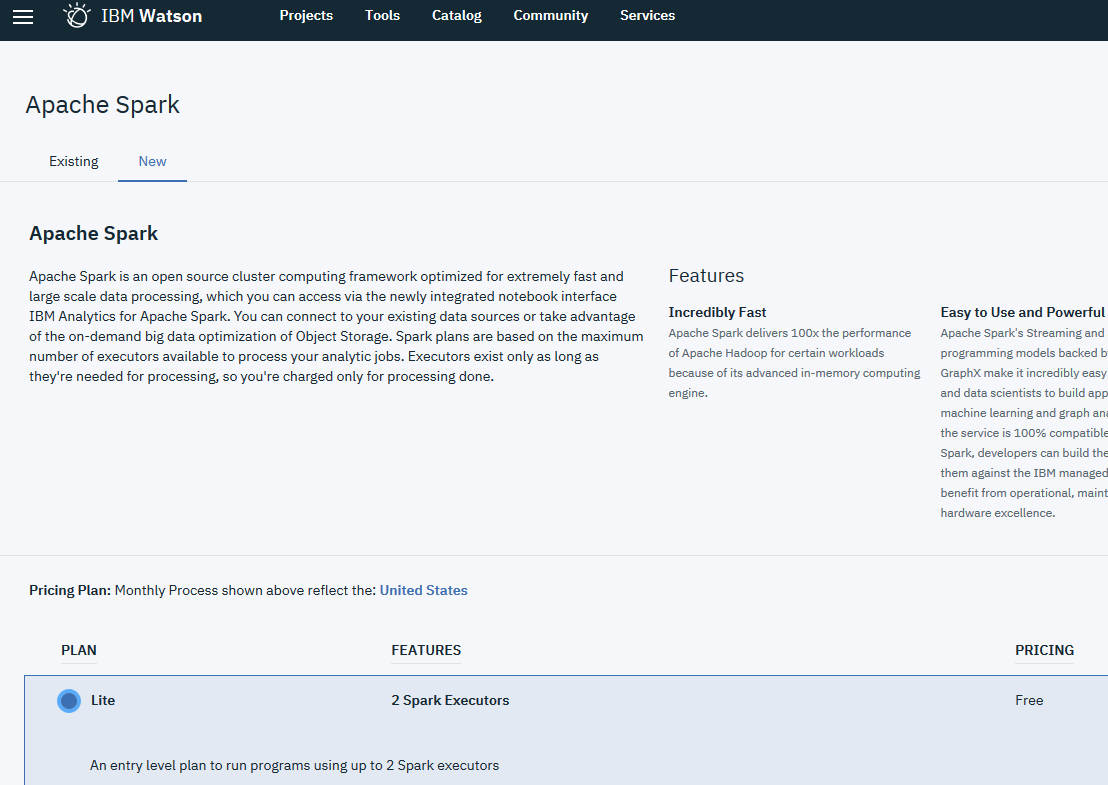
Abbiamo bisogno di attivare un servizio Apache Spark su cui fare lavorare il notebook. Dal menu di progetto scegli di aprire la pagina dei settings:



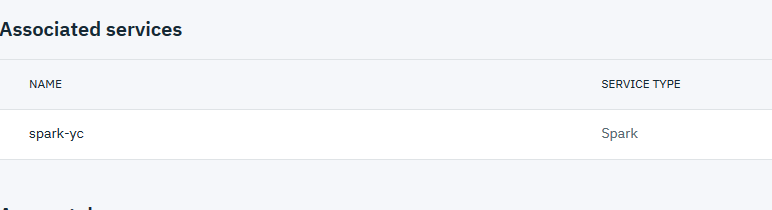
Nella pagina settings cerca la sezione Associated services, apri il menu a tendina e scegli il servizio Spark:



Si apre la finestra del servizio Apache Spark disponibile nel catalogo di IBM Cloud. Seleziona il piano Lite e procedi con **create** per instanziare il servizi. Il piano lite attiva due Spark executor.



A questo punto il servizio Spark compare tra i servizi associati al progetto.



**Lavoriamo con i Notebook**

## Creazione di un Notebook in Watson Studio

Un Notebook e’ un ambiente web-based di tipo interattivo. L’ambiente e’ organizzato in celle che possono contenere codice sorgente oppure testo descrittivo on formato markdown. Il codice contenuto nella singola cella puo’ essere eseguito e l’output viene mostrato a video immediatamente sotto alla cella stessa.

Quindi e’ possibile eseguire celle il cui risultato e’ la produzione di plot che vengono poi immediatamente visualizzati.

All’interno di Watson Studio e’ possibile scrivere Notebook nei linguaggi Scala, Python e R ed e’ possibile utilizzare framework come SparkMLlib, Scikit-Learn, TensorFlow e Keras. Qui trovi la lista completa:

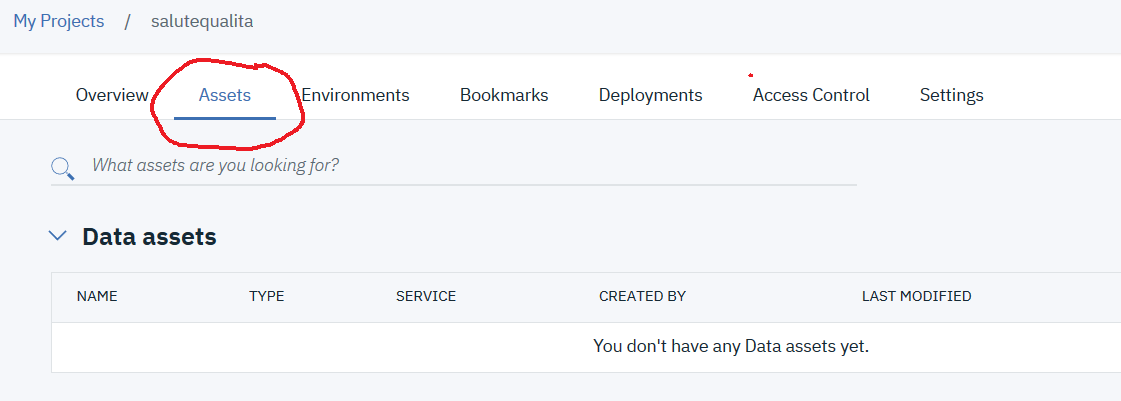
<https://datascience.ibm.com/docs/content/analyze-data/pm_service_supported_frameworks.html?context=analytics>

E’ possibile avere piu’ Notebook aperti contemporaneamente lavorando su finestre browser distinte. E’ possibile costruire un ambiente collaborativo a livello di progetto: i membri del team sono invitati a fare parte del progetto stesso e a condividerne non solo il codice sorgente ma anche gli output prodotti.

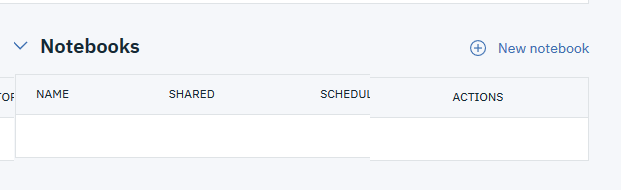
Un nuovo Notebook puo’ essere creato partendo da zero e creando tutte le sue celle, oppure caricando il Notebook da file o da URL.

Watson Studio include una grande vastita’ di librerie e package. Per installare una nuova libreria e’ possibile eseguirne l’installazione da cella di tipo Code usando le istruzioni standard del linguaggio in uso.

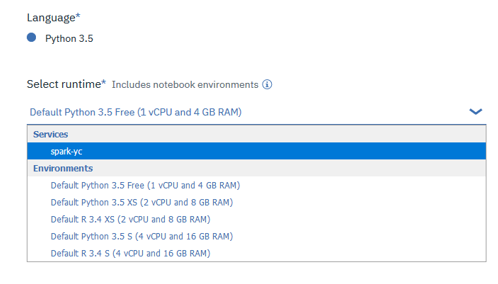
Dal menu di progetto scegli di aprire la pagina degli asset:



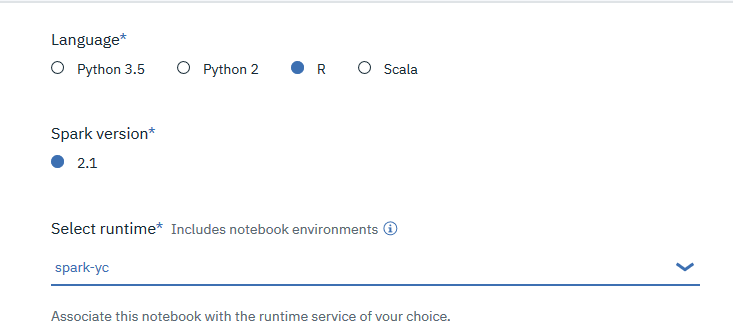
Vai nella sezione notebook e scegli di creare un new notebook:



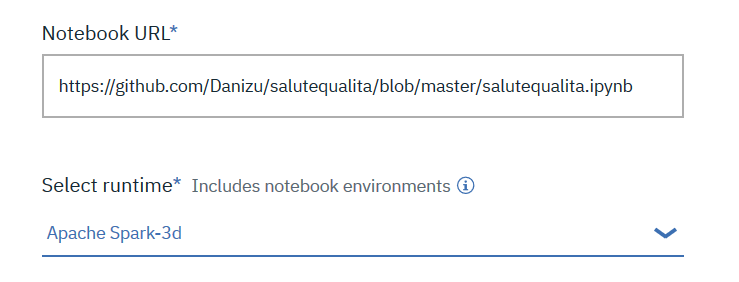
Vai nella sezione runtime e cambia il runtime che viene presentato scegliendo il servizio Spark appena creato:



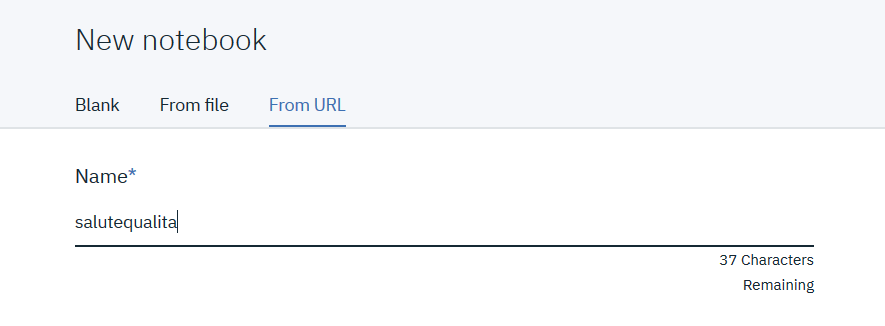
A questo punto cambia il linguaggio di programmazione e scegli R:



Il nuovo Notebook va creato a partire da URL. Quindi scegli **From URL** e nella box a fianco Notebook URL specifica questo URL <https://github.com/Danizu/RomaThink2018/blob/master/salutequalita4lab.ipynb>



Crea il nuovo notebook dando un nome:

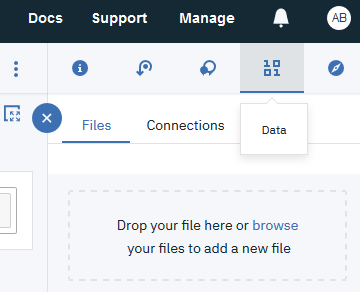


Procedi con **create notebook.**

Nel Notebook leggi con attenzione le celle di testo con le spiegazioni.

## Import del dataset nel Notebook in Watson Studio

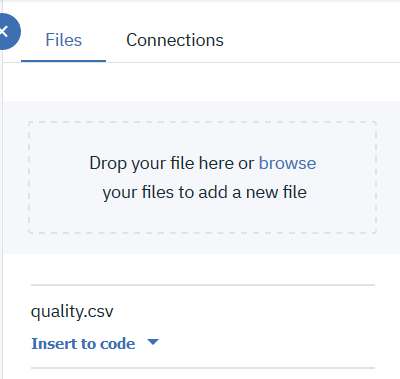
Si apre il notebook e compare la prima cella del notebook. La cella e’ vuota al momento. Carichiamo il dataset quality.csv. Nel menu in alto a destra scegliete di aprire la sezione dati cliccando sull’icona data:



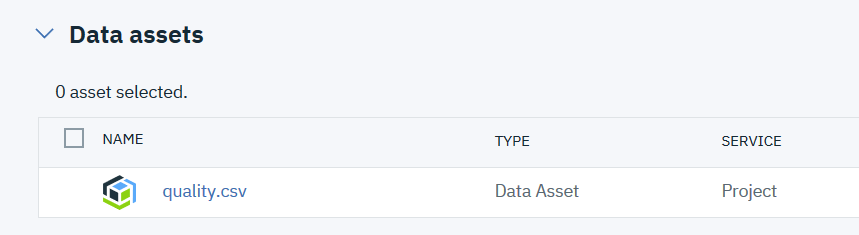
Seguendo le istruzioni, porta il file sull’object storage con drag and drop oppure facendo browse del disco. Il file puo essere scaricato da questo link:

<https://github.com/Danizu/salutequalita>

Per caricarlo nel progetto vai alla sezione dati e fai load del file:



Quando viene completato il caricamento del file nel progetto, il file compare anche nella sezione dei **data asset**:

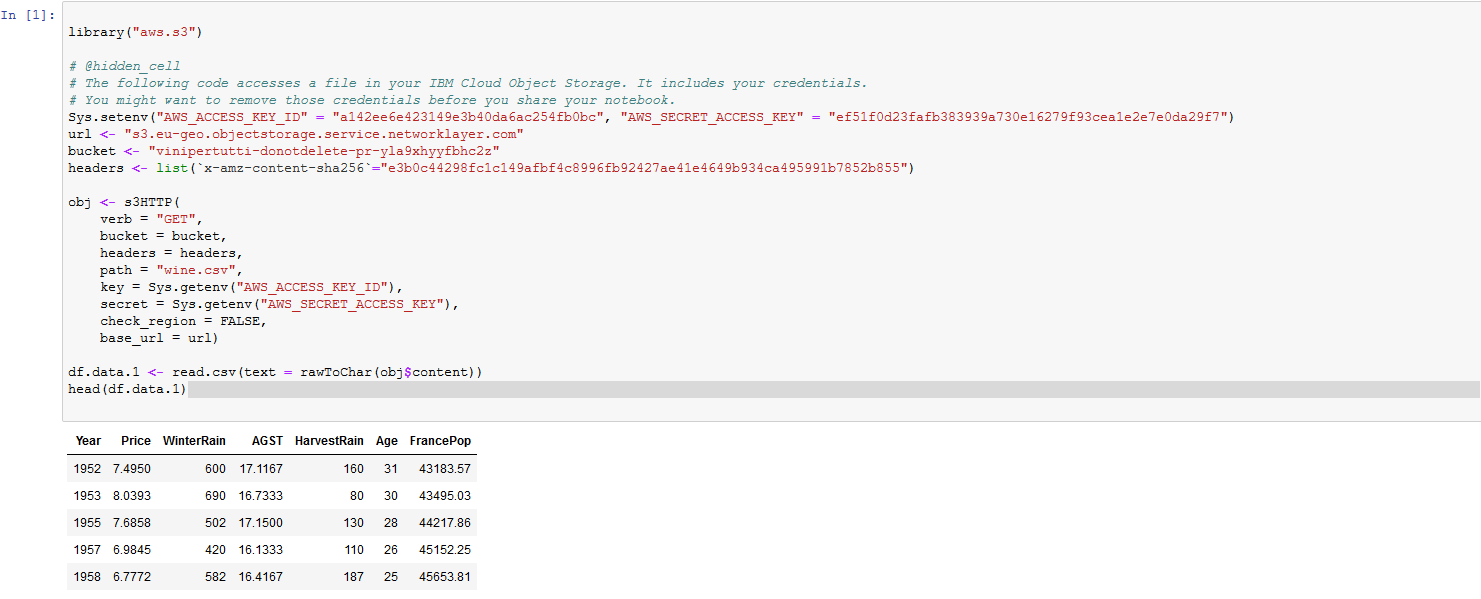


Posizionati con il cursore all’interno della prima cella di codice che contiene il commento : **# The code was removed by DSX for sharing**.

Apri il menu **insert to code** e scegli **Insert R DataFrame.**

Watson Studio genera automaticamente una cella di codice in linguaggio R che contiene le variabili ambiente per accedere e leggere l’object storage, effettua la lettura del file quality.csv che abbiamo caricato sull’object storage e carica il contenuto del file in un dataset. Fai attenzione alle istruzioni contenute nel testo del Notebook. In particolare **rinomina il nome del dataset da df.data.1 in quality.**

Con la combinazione di tasti CTRL-ENTER, la cella viene eseguita:

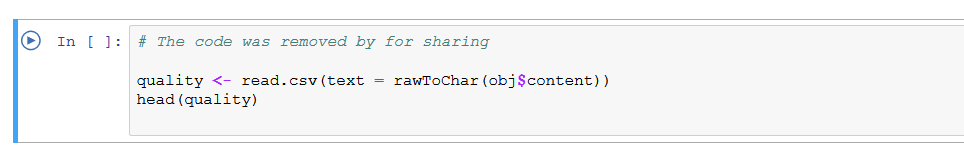


## 

## Esecuzione nel Notebook in Watson Studio

Leggi con attenzione il contenuto del Notebook ed esegui, in modo interattivo, il codice contenuto nelle celle.

Per eseguire una cella di codice, posizionati con il mouse sulla cella ed esegui CTRL-ENTER oppure fai click sulla freccia a sinistra accanto alla cella:



Il notebook contiene delle celle incomplete da usare come esercizio. Esegui le celle completando con il codice mancante.

## Reference

IBM Cloud: <https://www.ibm.com/cloud/>

IBM Watson Studio: <https://datascience.ibm.com/>

IBM Watson Machine Learning: <https://www.ibm.com/support/knowledgecenter/DSXDOC/analyze-data/ml-overview.html>

IBM Apache Spark: <https://www.ibm.com/cloud/spark>

Cognitive class: <https://cognitiveclass.ai/>

Sommelier Lab: <https://github.com/Danizu/siamosommelier>

IBM Cloud Object Storage (COS): <https://www.ibm.com/cloud/object-storage>

# D. Bertsimas Algorithmic Prediction of Health-Care Costs:

# [https://dl.acm.org/citation.cfm?id=1528038](https://dl.acm.org/citation.cfm?id=1528038%20)

Acknowledgements and Disclaimers

Copyright © 2017 by International Business Machines Corporation (IBM). No part of this document may be reproduced or transmitted in any form without written permission from IBM.

**U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed “as is” without any warranty, either express or implied. In no event shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity**. IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply.”

**Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.** Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer’s responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM’s products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a particular, purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right. IBM, the IBM logo, ibm.com, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services®, Global Technology Services®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli® Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).