Cloud Pak for Data JumpStart Training

Welcome to the training.

Please try and login to the platform, https://cp4d.apps.ocp.dmp.qa.bhp.com/ using your BHP user id and Windows password. *Let me know immediately if you are unable to login.*

Download today's training material from,

https://github.com/randyphoaibm/cpd-jumpstart/raw/main/CPD_JumpStart_Training.pdf

https://raw.githubusercontent.com/randyphoaibm/cpd-jumpstart/main/german_credit.csv

Cloud Pak for Data JumpStart Training

Randy Phoa
Senior Data Scientist
IBM Data Science Elite Team

Agenda

1. Introduction to Cloud Pak for Data
2. Projects
3. Prepare data
4. Build models
5. Deploy models
6. Resources

1. Introduction to Cloud Pak for Data

Cloud Pak for Data
Unified, modular, deployable anywhere

App Developers | Business Analysts | Data Engineers | Data Stewards | Data Scientists | Business Users

Integrated User Experience

Extensible: APIs, partner ecosystem, accelerators, and solutions

Collect
- Data virtualization
- SQL and NoSQL databases
- Event ingestion
- Streaming Analytics
- Apache Spark

Organize
- Data transformation
- Data quality and classification
- Policies and rules
- Data cataloging
- Self-service discovery and search

Analyze
- Data science and visualization
- AutoAI
- Model trust and explainability
- Model optimization

Infuse
- Business reporting and visualization
- Financial planning and analysis
- Cloud native AI services
- RegTech and Financial Crimes Insight

Cloud Pak core services
Security, Administration, Operations

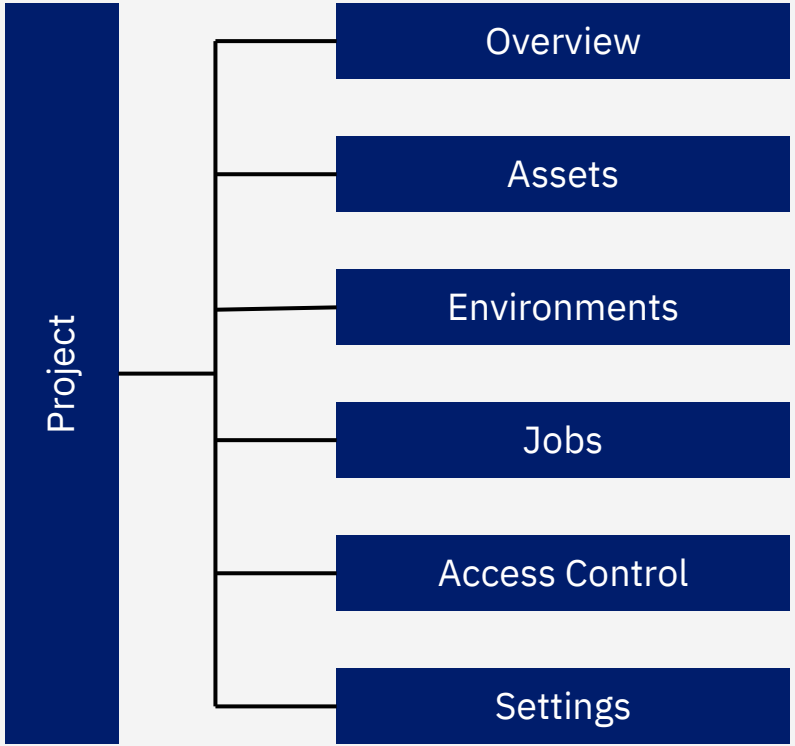
Red Hat OpenShift

IBM Cloud | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

2. Projects

- Project structure
- Git Setup
- Create a project
- Add collaborators
- Custom environments

Project structure



Meta data of the project

A list of all the assets, including data, notebooks, etc.

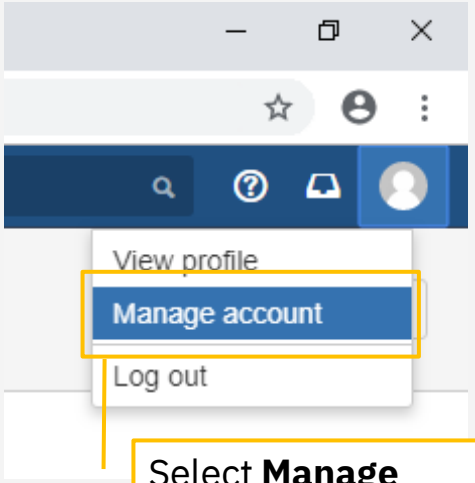
Default as well as manually defined environment

Jobs

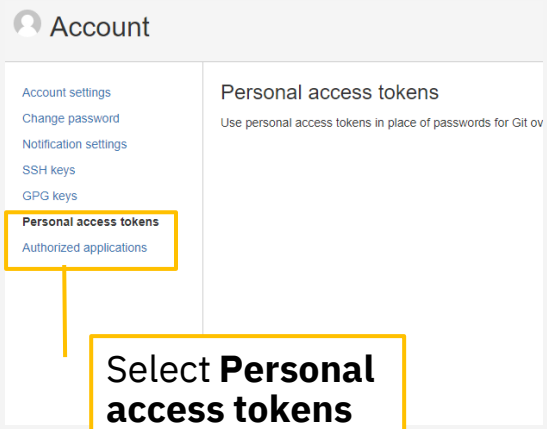
Collaborators with different access level

Deployment space as well as integration with Git

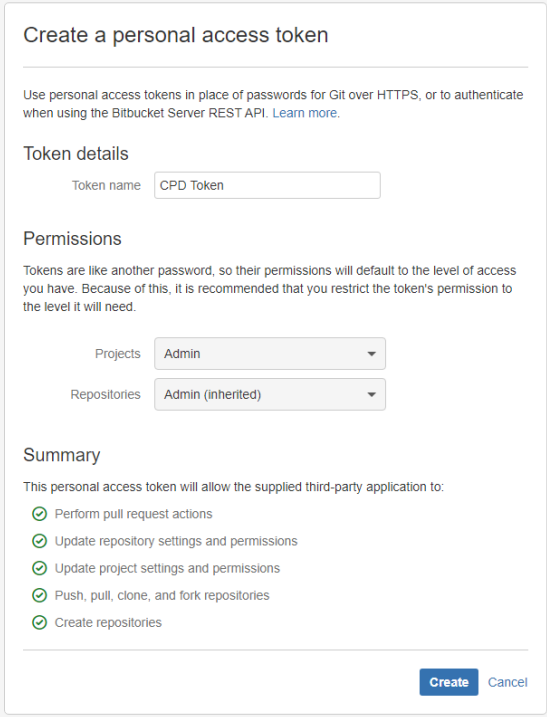
Git Setup – Create access token from BitBucket

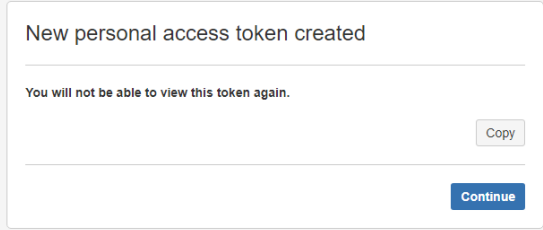


Select **Manage account** from menu

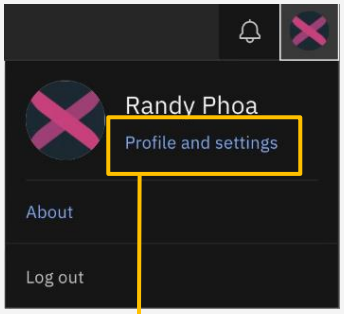


Select **Personal access tokens**

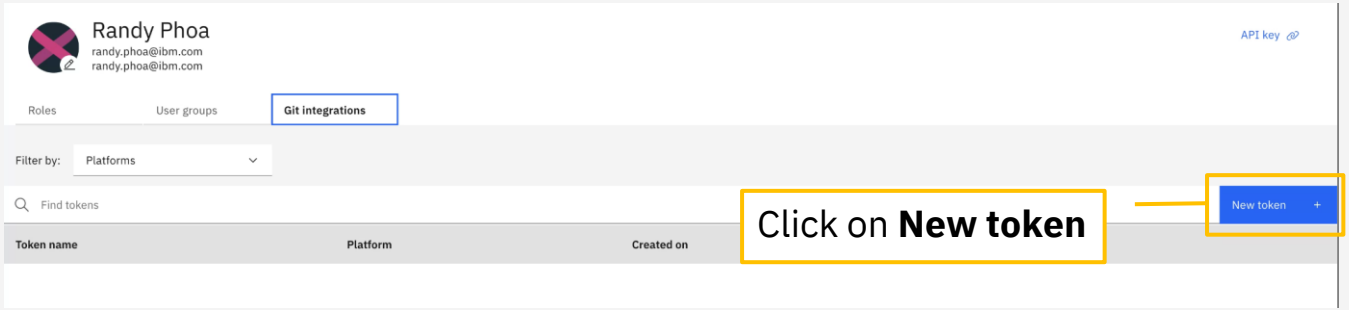




Git Setup – Add access token to Cloud Pak for Data

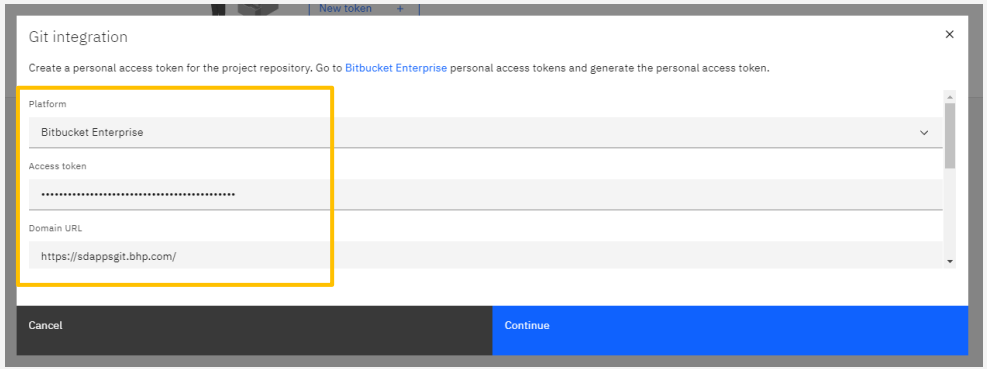


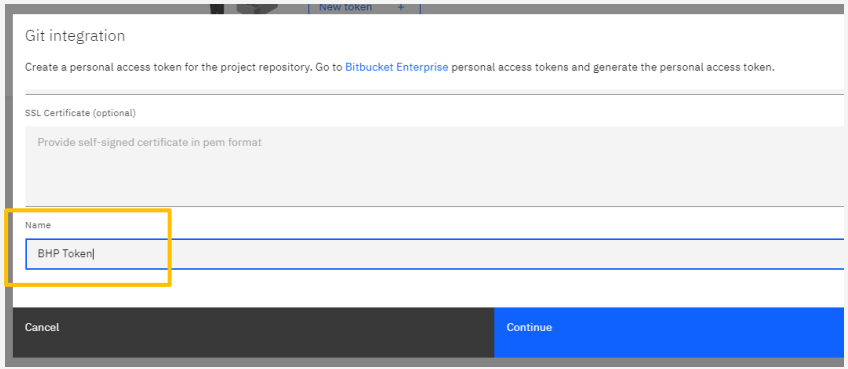
Select **Profile and settings** from menu

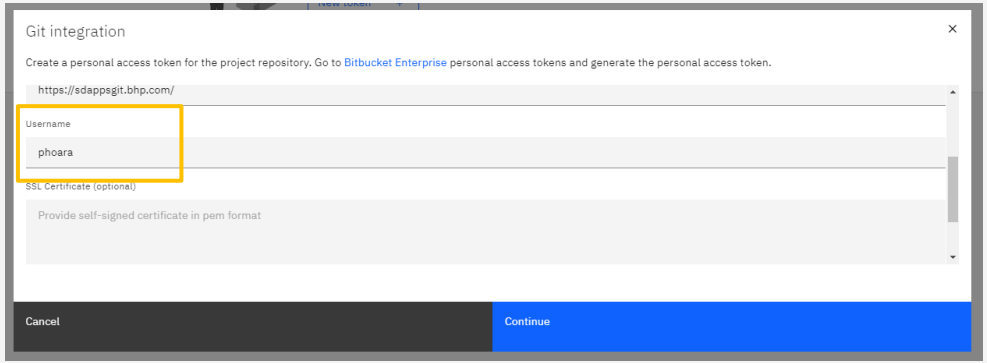


Click on **New token**

Git Setup – Add access token to Cloud Pak for Data







Fill in required details.
Click **Create** when done.
Remember to save your access token securely.

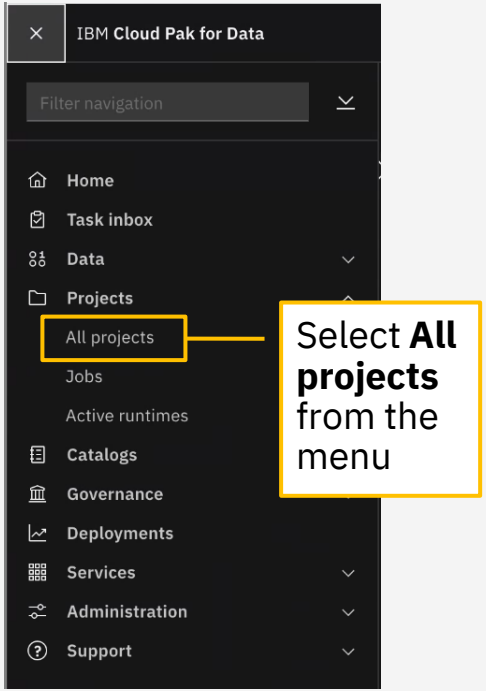
Exercise

- Create your BitBucket access token
- Add your access token to Cloud Pak for Data

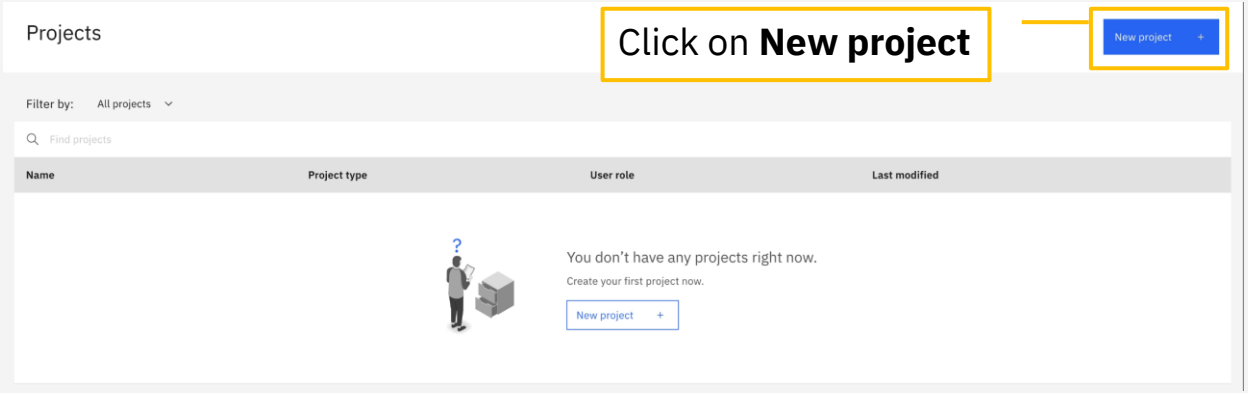
Lesson

- Ensure you know how to create a BitBucket access token and how to add it to Cloud Pak for Data

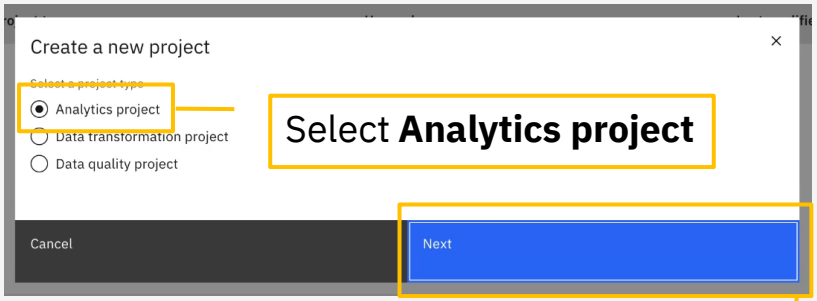
Create a project



Select **All projects** from the menu

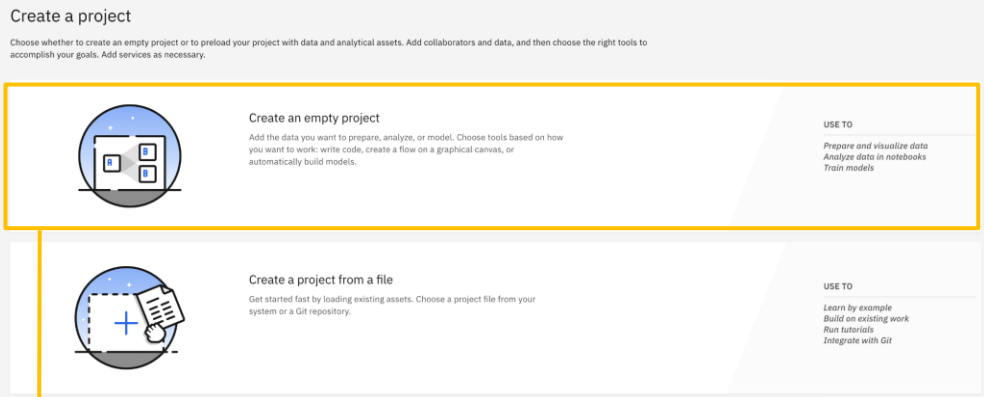


Create a project



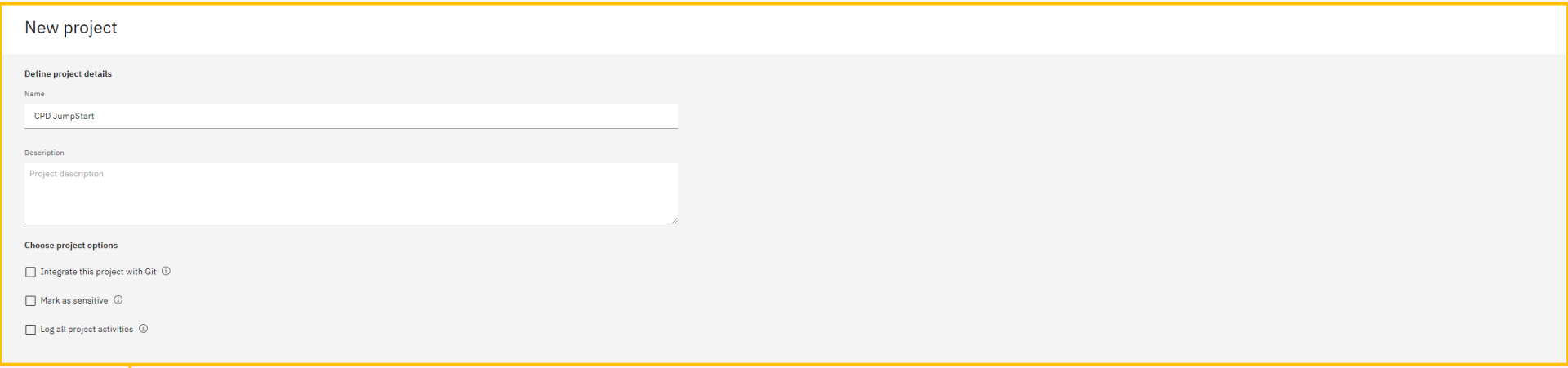
Select **Analytics project**

Click on **Next**



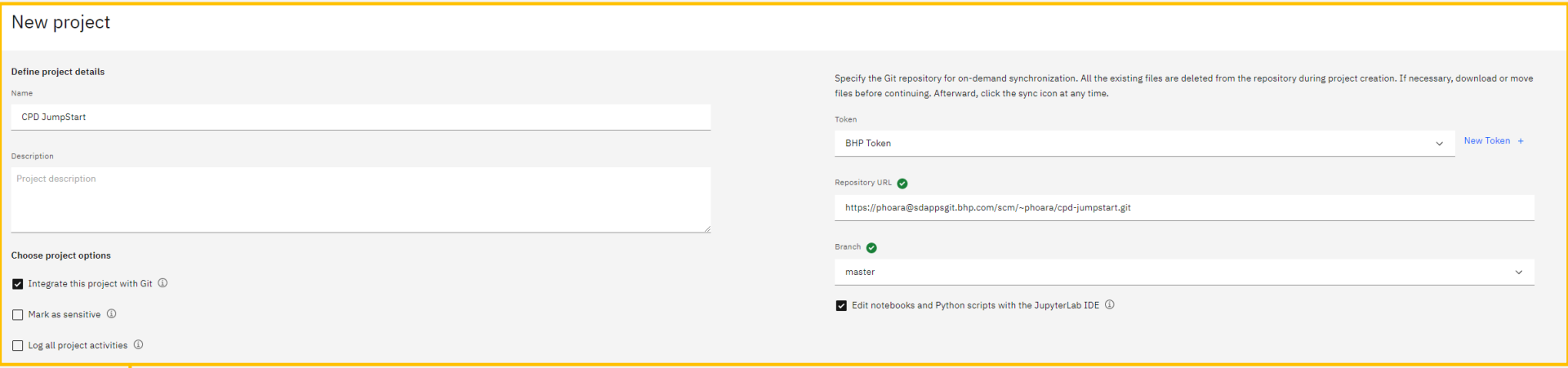
Select **Create an empty project**

Create a project – without Git integration



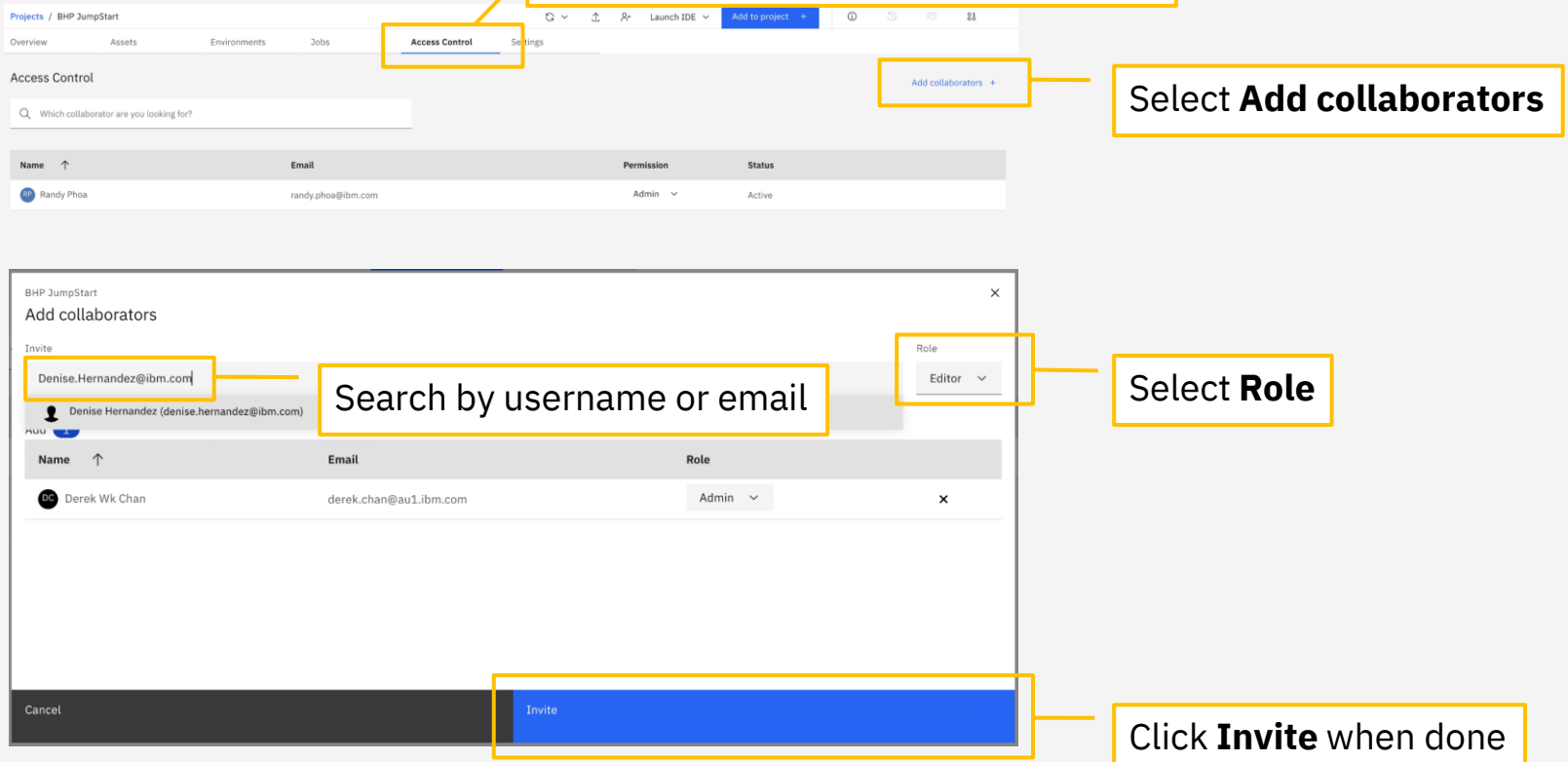
Fill in the required details

Create a project – with Git integration



Fill in the required details

Add collaborators



Roles

Role	Permissions	Services that contribute permissions	Service that creates the role
Administrator	– Administer platform – Create service instances	Cloud Pak for Data control plane	Cloud Pak for Data control plane
	– Integrate and transform data	DataStage® Edition	
	– Analyze data quality – Discover assets – Import metadata – Manage catalogs – Access governance artifacts – Manage governance categories – Manage governance workflows – Manage information assets – Manage data protection rules	Watson™ Knowledge Catalog	
Business Analyst	– Access information assets view – View data quality	Watson Knowledge Catalog	Watson Knowledge Catalog
Data Engineer	– Create service instances	Cloud Pak for Data control plane	DataStage Edition or Watson Knowledge Catalog
	– Integrate and transform data	DataStage Edition	
	– Access catalogs – Discover assets – Import metadata – Access governance artifacts – Manage information assets – View data quality	Watson Knowledge Catalog	
Data Quality Analyst	– Access catalogs – Analyze data quality – Discover assets – Import metadata – Access governance artifacts – Manage information assets	Watson Knowledge Catalog	Watson Knowledge Catalog
Data Scientist	– Access catalogs	Watson Knowledge Catalog	Watson Knowledge Catalog

Role	Permissions	Services that contribute permissions	Service that creates the role
Data Steward	– Access catalogs – Discover assets – Import metadata – Access governance artifacts – Manage information assets – View data quality – Manage data protection rules	Watson Knowledge Catalog	Watson Knowledge Catalog
Developer	– Create service instances	Cloud Pak for Data control plane	Watson Knowledge Catalog
	– Access catalogs	Watson Knowledge Catalog	
User	– Access assigned services	Cloud Pak for Data control plane	Cloud Pak for Data control plane

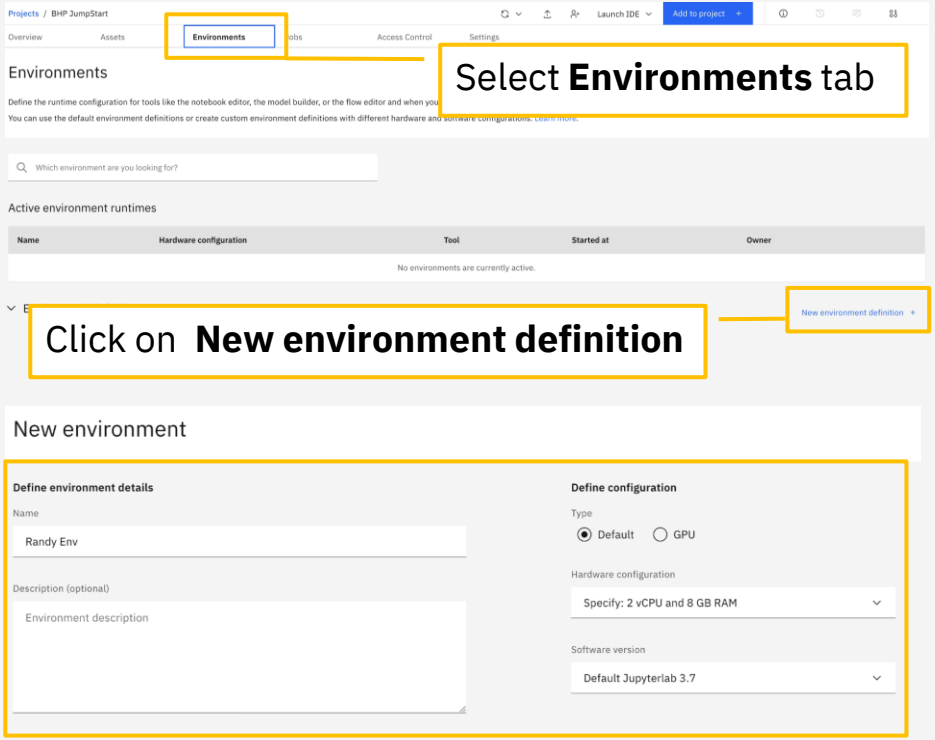
Exercise

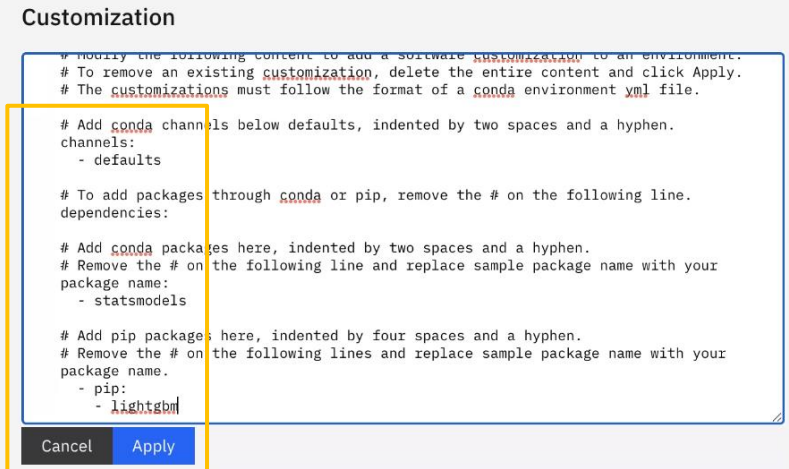
- Create a project without Git integration
- Add a few of your colleagues from your team/department
- Ensure they can see your project (and vice versa if they added you)

Lesson

- Ensure you know how to create a project and add collaborators

Custom environments

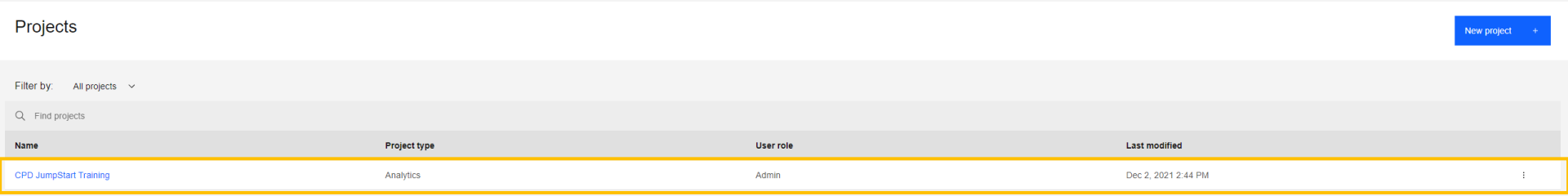




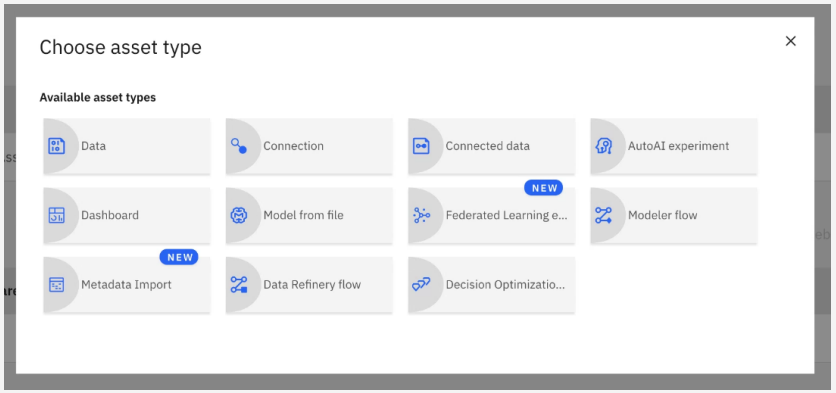
3. Prepare data

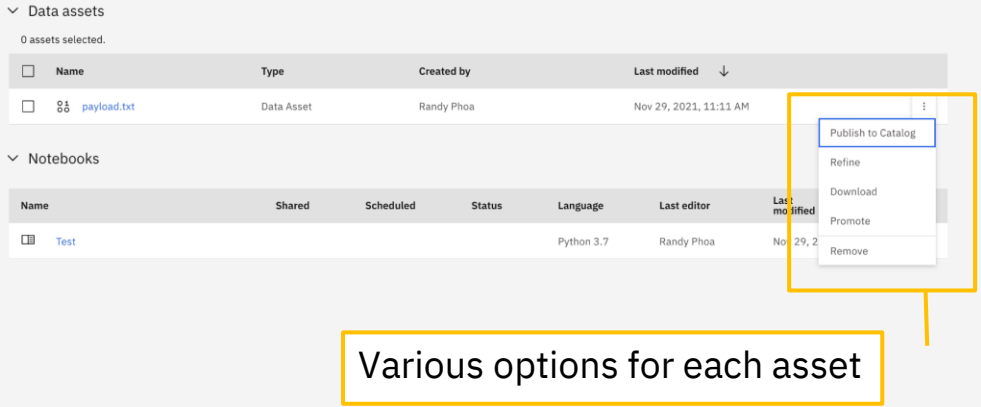
- Manage assets
- Upload files
- Create data connection

Open the **CPD JumpStart Training** project



Manage assets





Various options for each asset

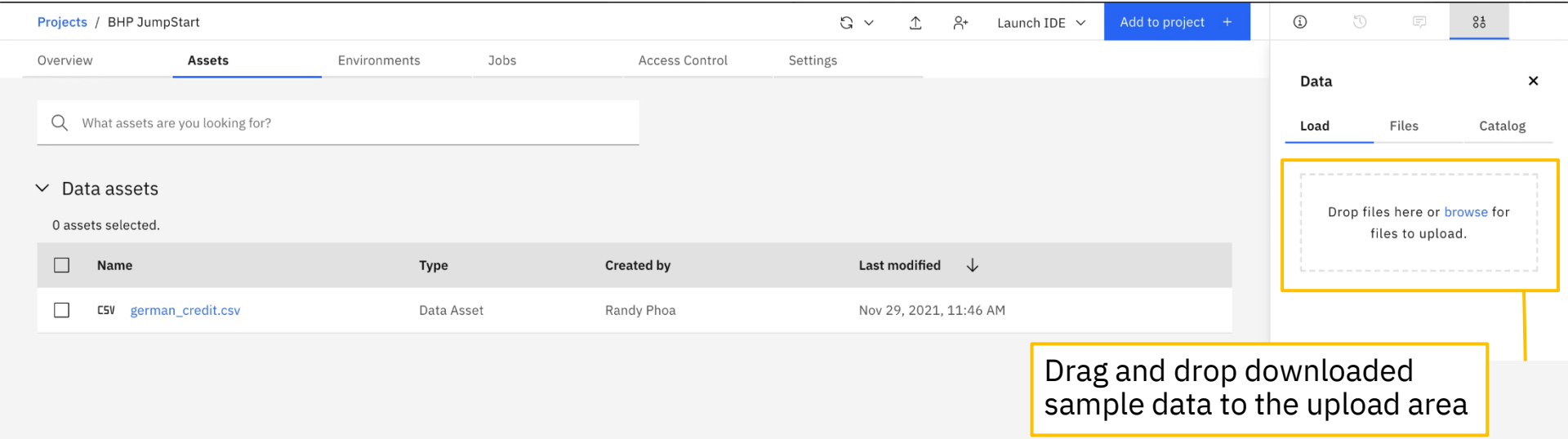
Download sample data

Access the below link and download the file to your local computer by clicking on File -> Save As.

https://raw.githubusercontent.com/randyphoaibm/cpd-jumpstart/main/german_credit.csv

Upload data

Local Computer -> Watson Studio



Drag and drop downloaded sample data to the upload area

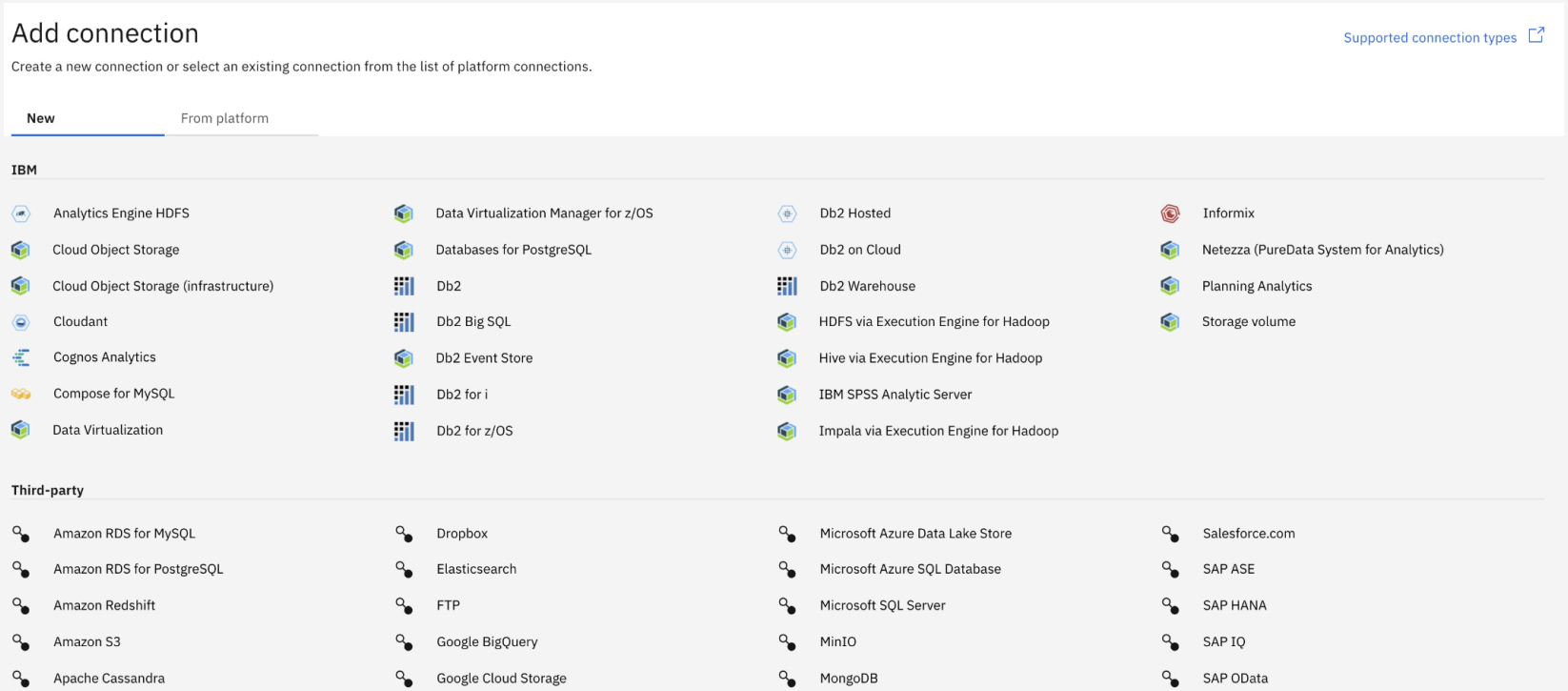
Exercise

- Download the sample data and add a suffix using your first name, last name and 4 random numbers, i.e, (german_credit_RandyP-8650.csv)
- Upload the file by dragging and dropping to the upload area
- Ensure that you see the file appear in Data Assets

Lesson

- Understand how files are uploaded and shared with collaborators in the project

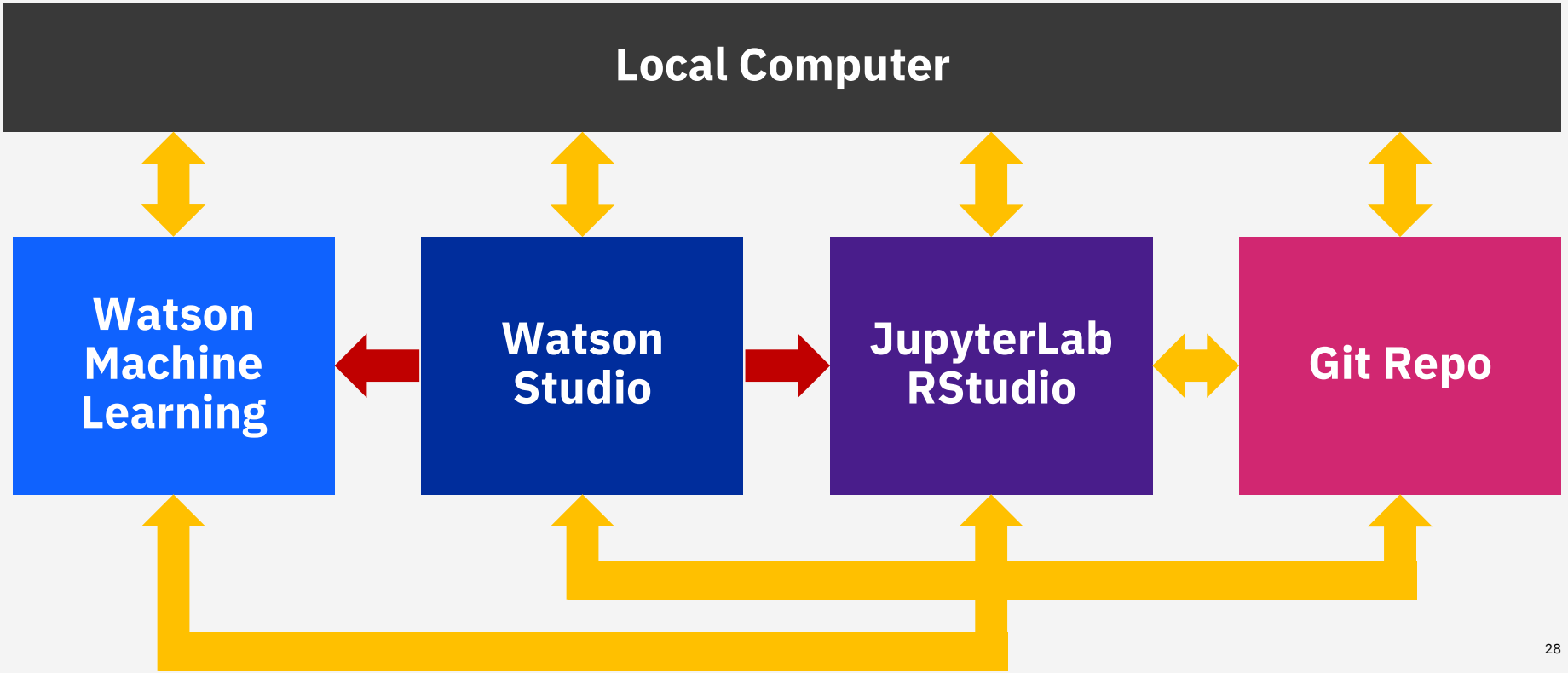
Create data connection



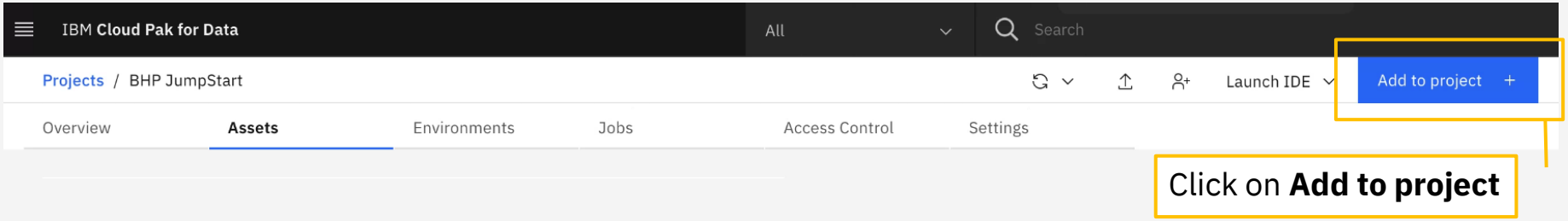
4. Develop models

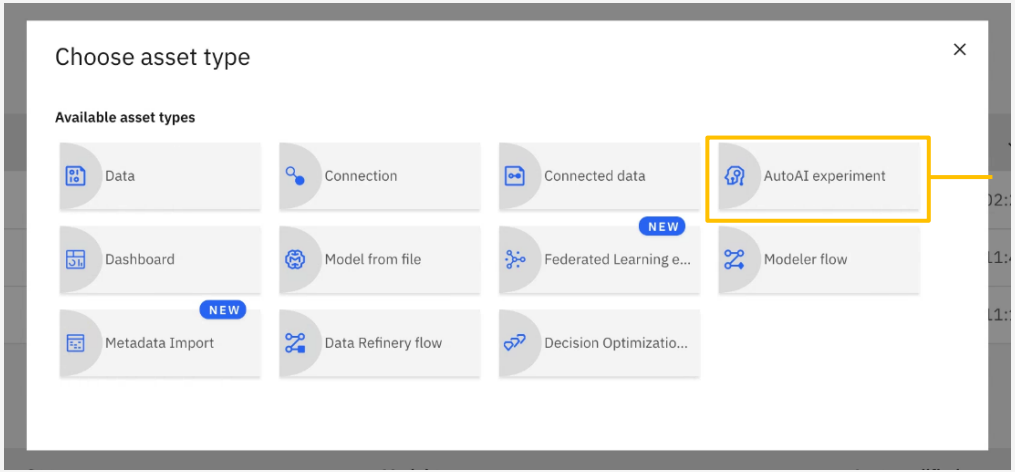
- AutoAI
- RStudio
- JupyterLab

Data flow among the various environments

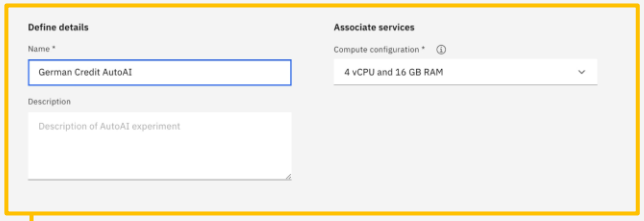


AutoAI – Create AutoAI experiment

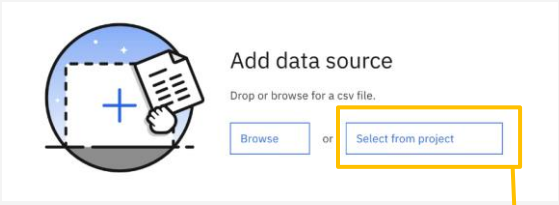




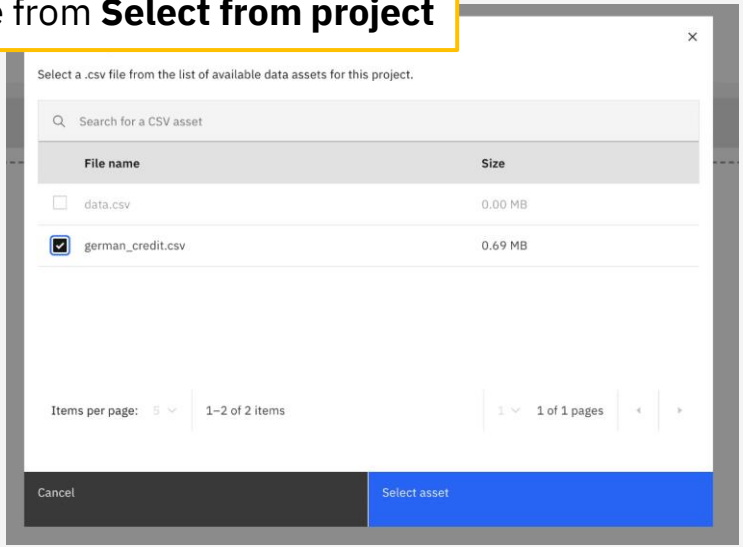
AutoAI – Configure



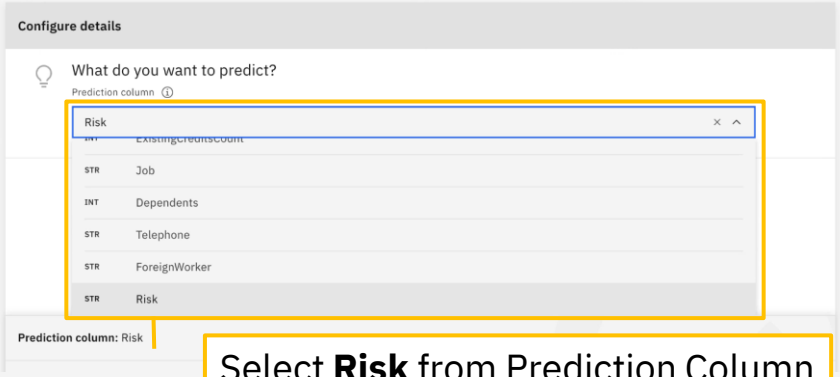
Fill in the required details



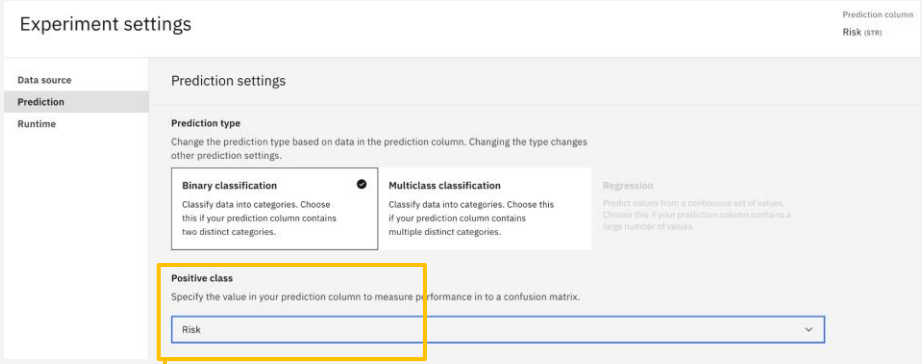
Add data source from **Select from project**



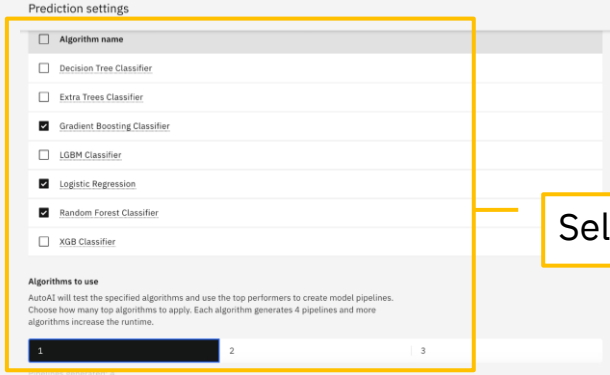
AutoAI – Settings



Select **Risk** from Prediction Column

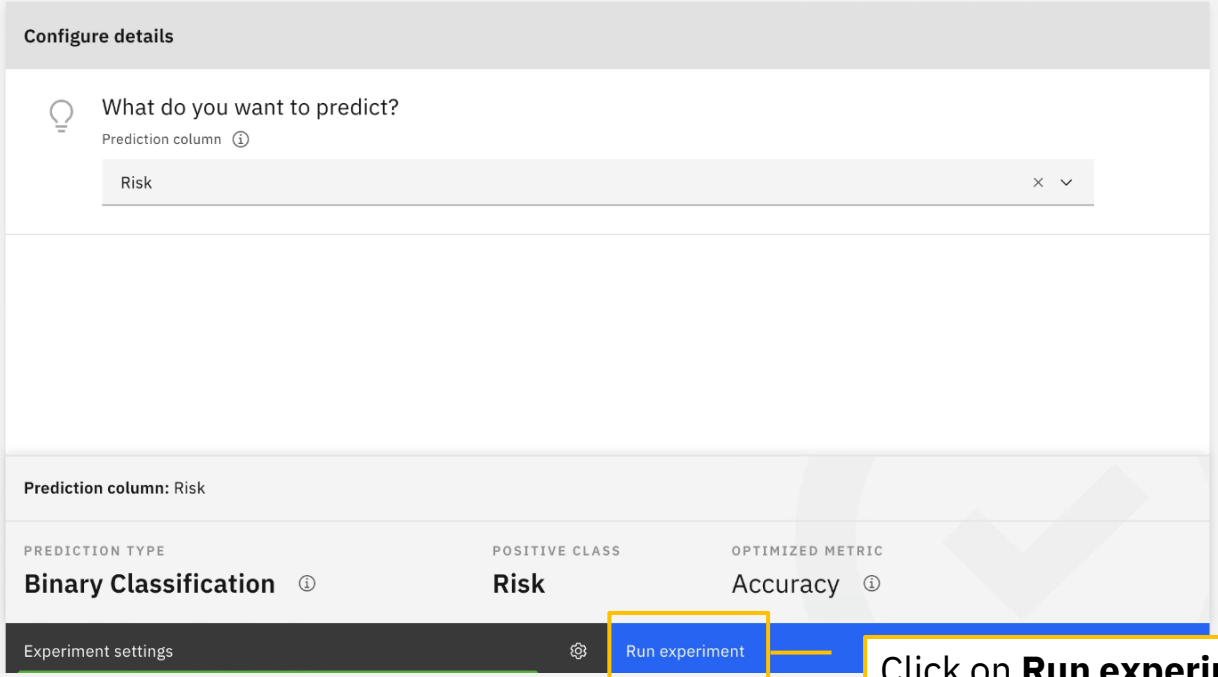


From **experiment settings**, select the **Prediction** tab and select **Risk** as the **Positive Class**



Select **Algorithms**

AutoAI – Run experiment



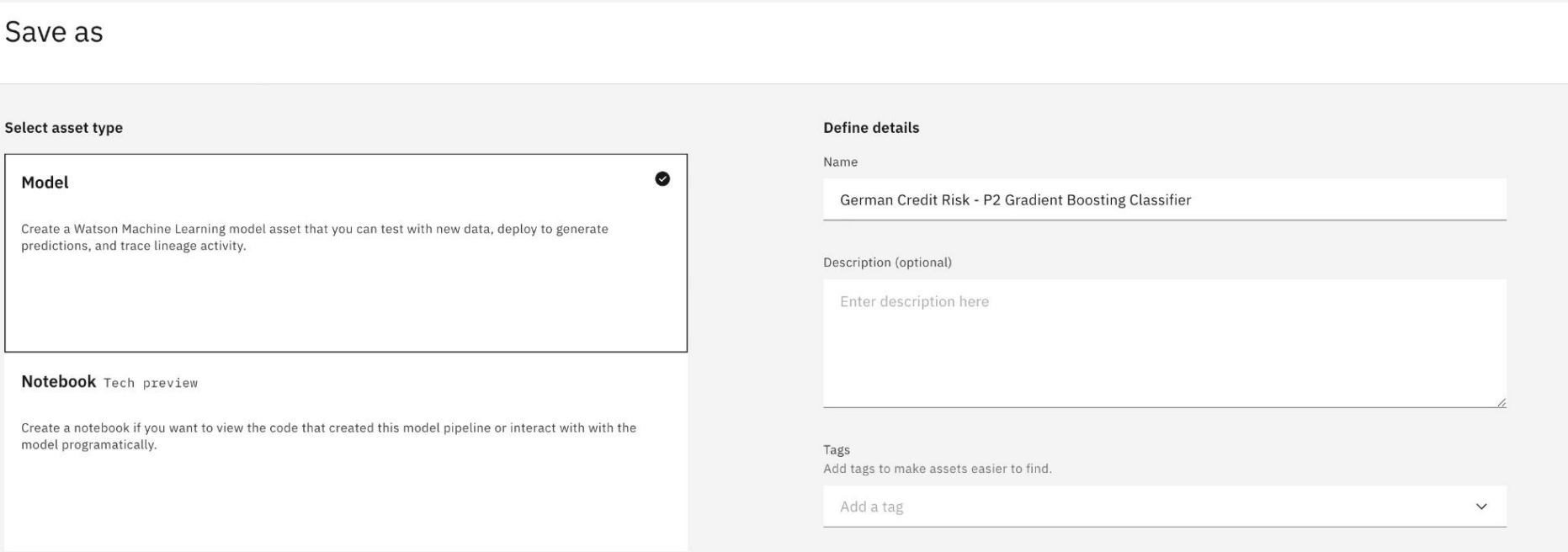
Click on **Run experiment**

AutoAI – Results

Pipeline leaderboard

Rank ↑	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
★ 1	Pipeline 2	Gradient Boosting Classifier	**0.805**	HPO-1	00:00:36
2	Pipeline 3	Gradient Boosting Classifier	**0.805**	HPO-1 FE	00:04:52
3	Pipeline 4	Gradient Boosting Classifier	**0.805**	HPO-1 FE HPO-2	00:01:16
4	Pipeline 1	Gradient Boosting Classifier	**0.799**	*None*	00:00:07
5	Pipeline 8	XGB Classifier	**0.789**	HPO-1 FE HPO-2	00:01:22
6	Pipeline 7	XGB Classifier	**0.788**	HPO-1 FE	00:01:12
7	Pipeline 5	XGB Classifier	**0.783**	*None*	00:00:02

AutoAI – Save as



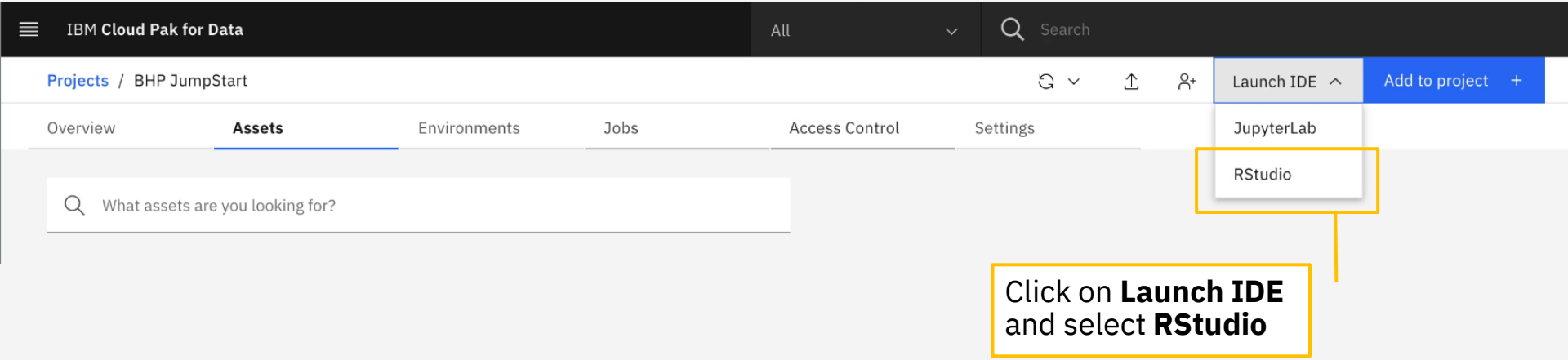
Exercise

- Create an AutoAI experiment using the sample data provided

Lesson

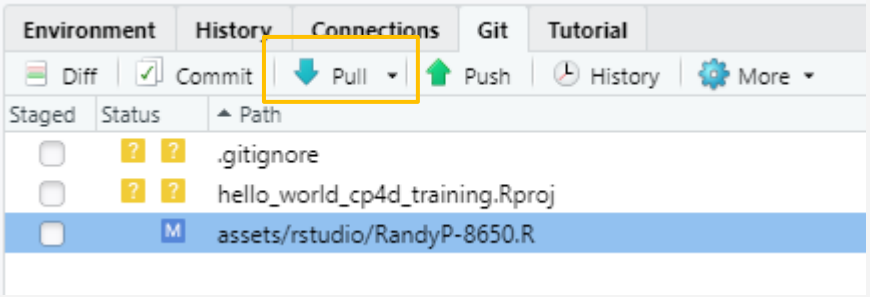
- Learn how to create and configure an AutoAI experiment

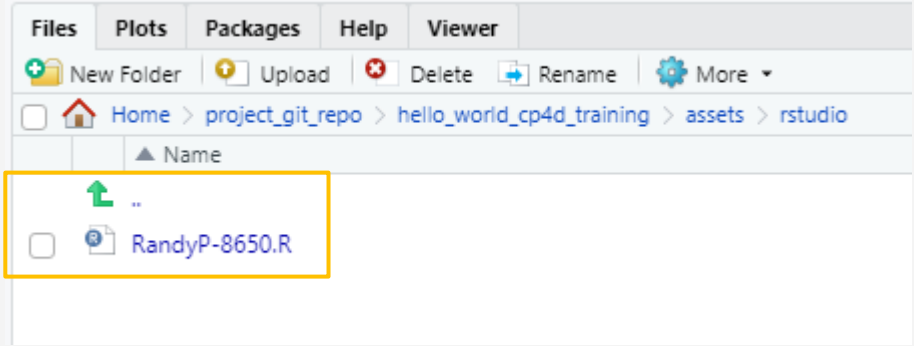
RStudio



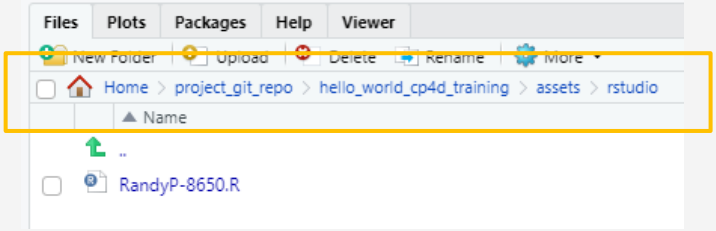
Click on **Launch IDE** and select **RStudio**

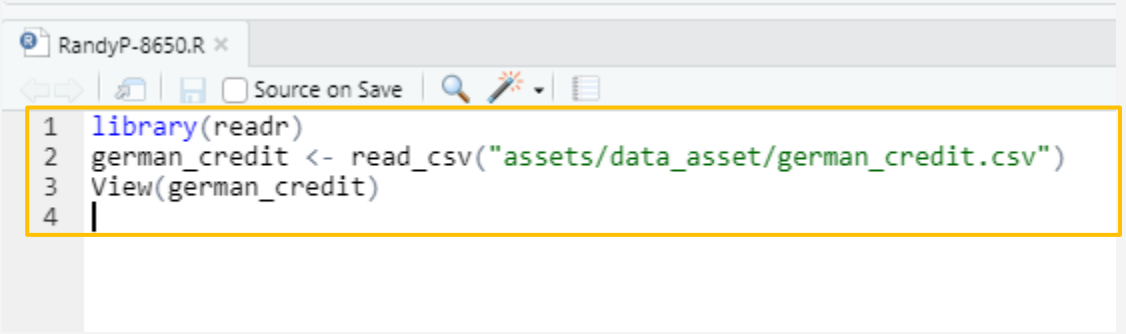
RStudio – Pull from repo



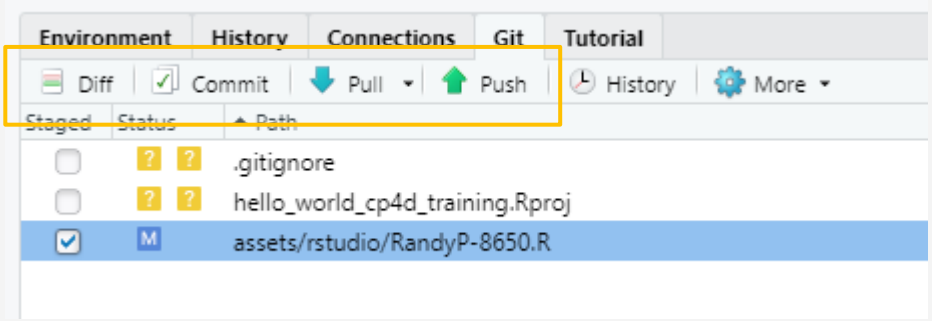


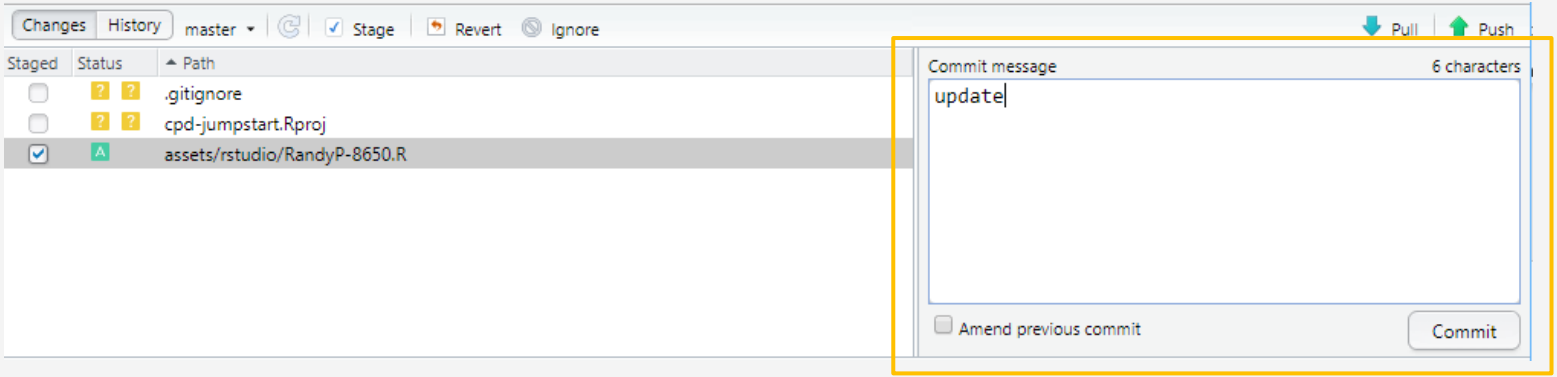
RStudio – Add script





RStudio – Push to repo





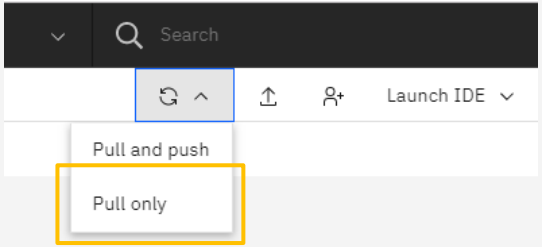
Exercise

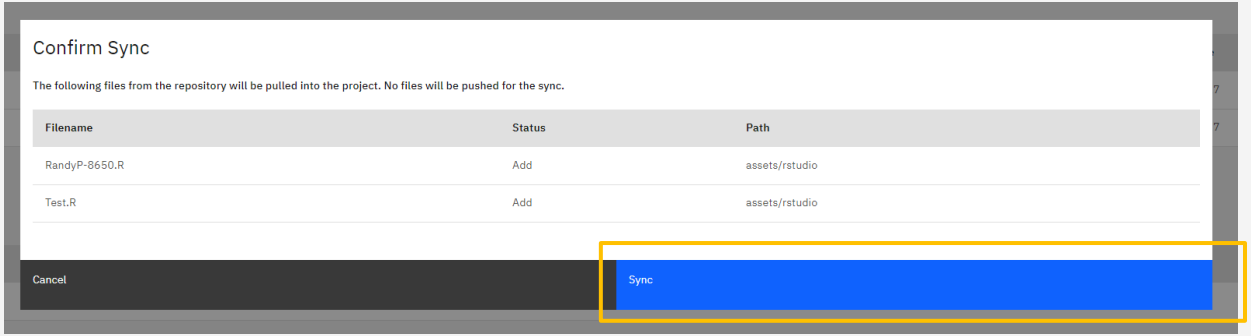
- Launch the RStudio IDE
- Pull from repo
- Add a R script
- Push to repo

Lesson

- Understand how the different file environments (Watson Studio, RStudio and Git repo work together)

Watson Studio – Check R Scripts

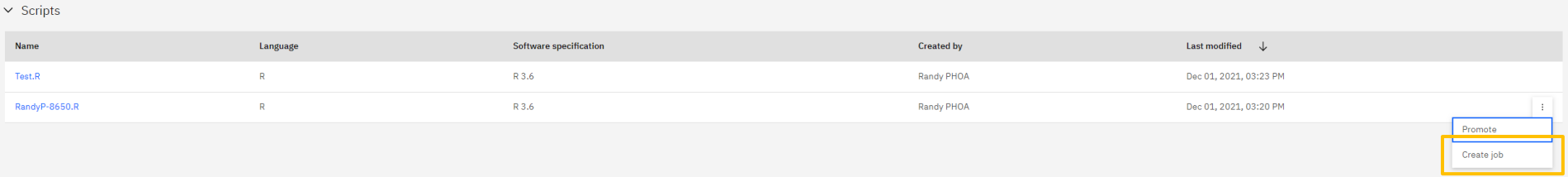


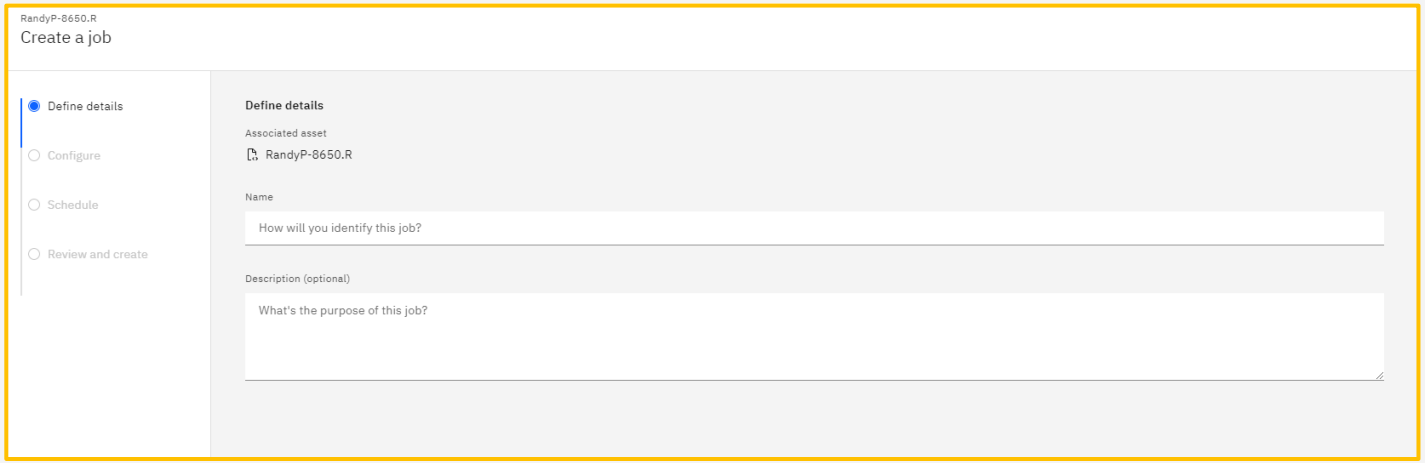


Scripts

Name	Language	Software specification	Created by	Last modified
Test.R	R	R 3.6	Randy PHOA	Dec 01, 2021, 03:23 PM
RandyP-8650.R	R	R 3.6	Randy PHOA	Dec 01, 2021, 03:20 PM

Run jobs in Watson Studio – R Scripts





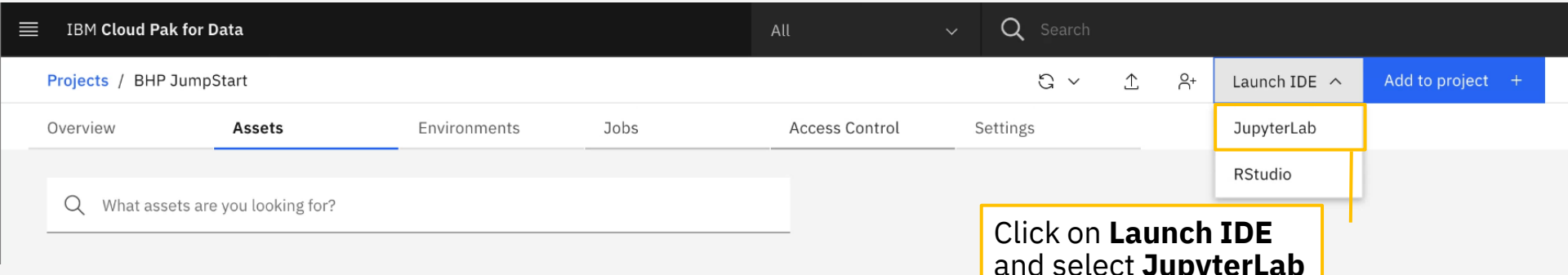
Exercise

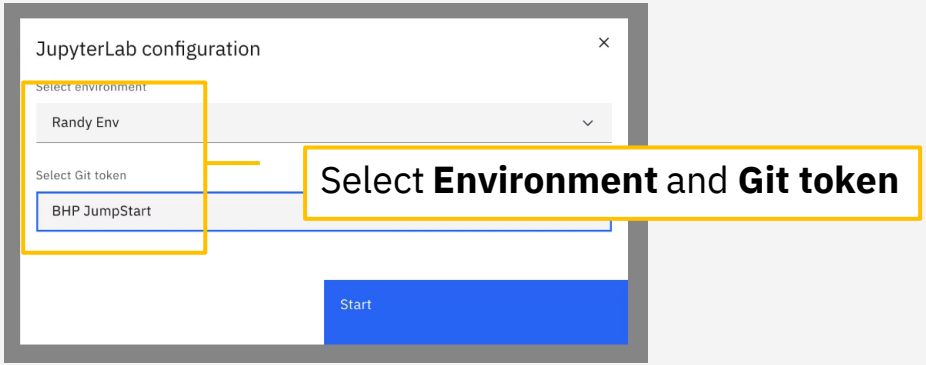
- Sync project with repo
- Ensure you can see your R Script
- Create a Job

Lesson

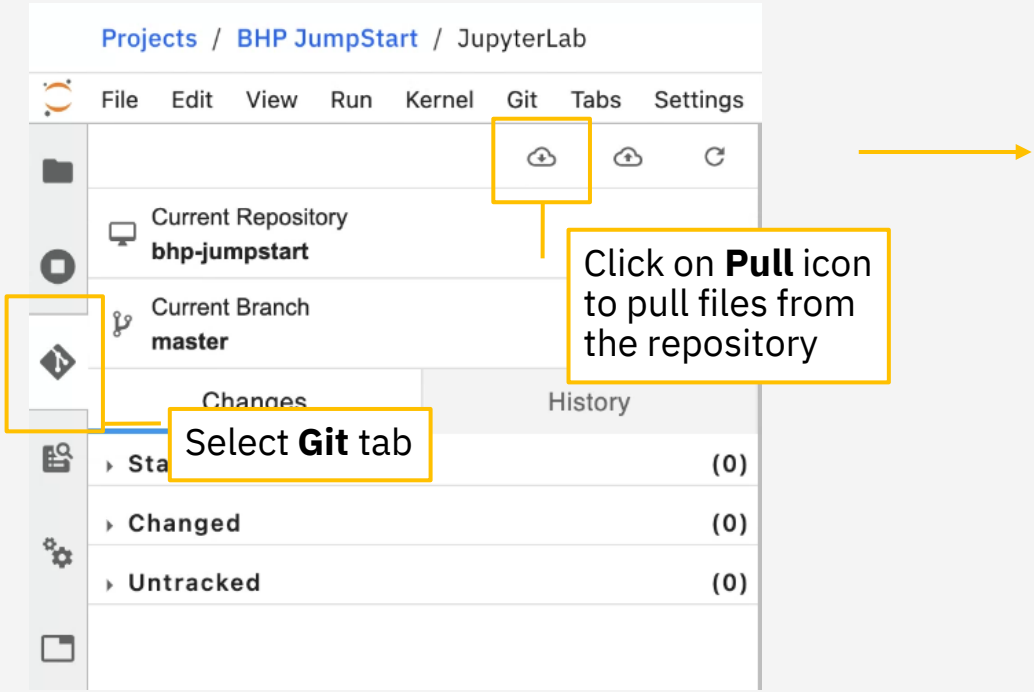
- Understand how the different file environments (Watson Studio, RStudio and Git repo work together)
- Learn how to create and configure a Job

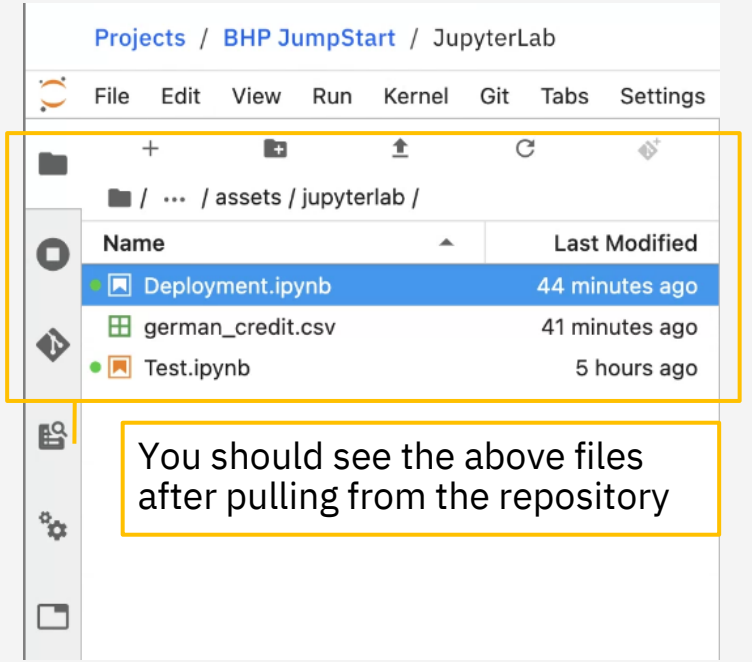
JupyterLab


IBM Cloud Pak for Data
All
Search
Projects / BHP JumpStart
Launch IDE
Add to project
Overview
Assets
Environments
Jobs
Access Control
Settings
JupyterLab
RStudio
What assets are you looking for?
Click on **Launch IDE** and select **JupyterLab**

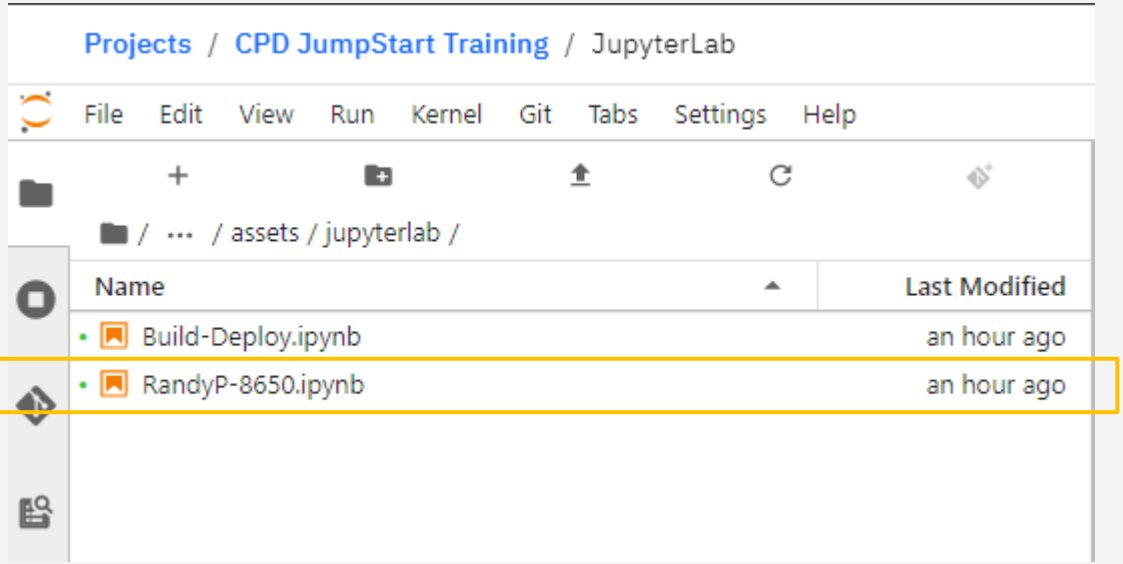

JupyterLab configuration
Select environment
Randy Env
Select Git token
BHP JumpStart
Start
Select **Environment** and **Git token**

JupyterLab – Pull from repo

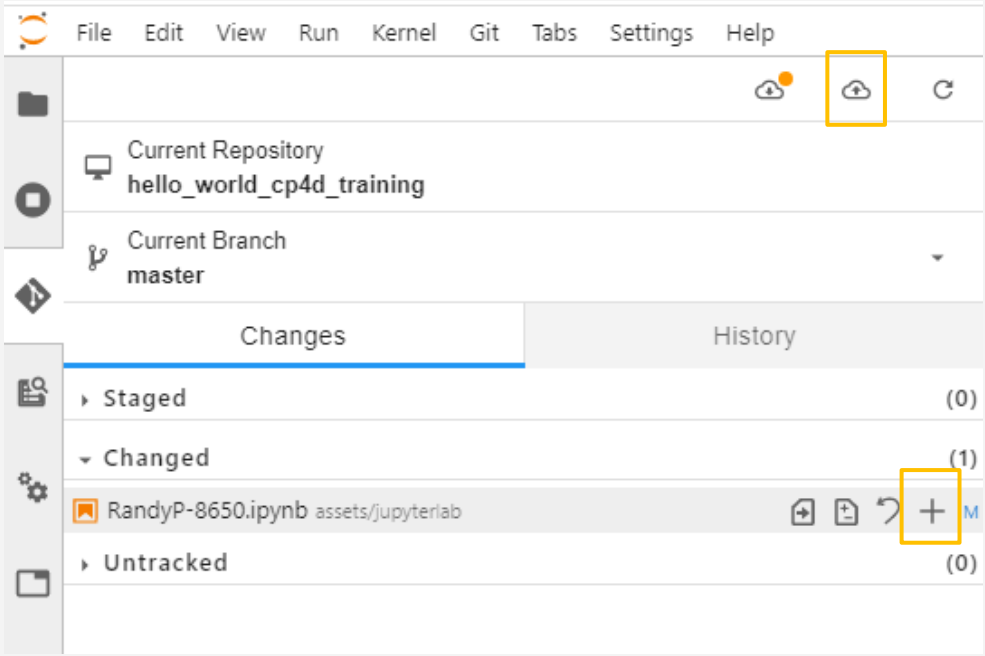




JupyterLab – Add notebook



JupyterLab – Push to repo



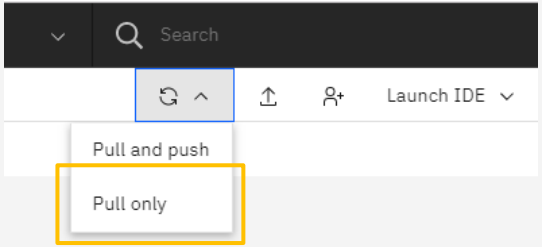
Exercise

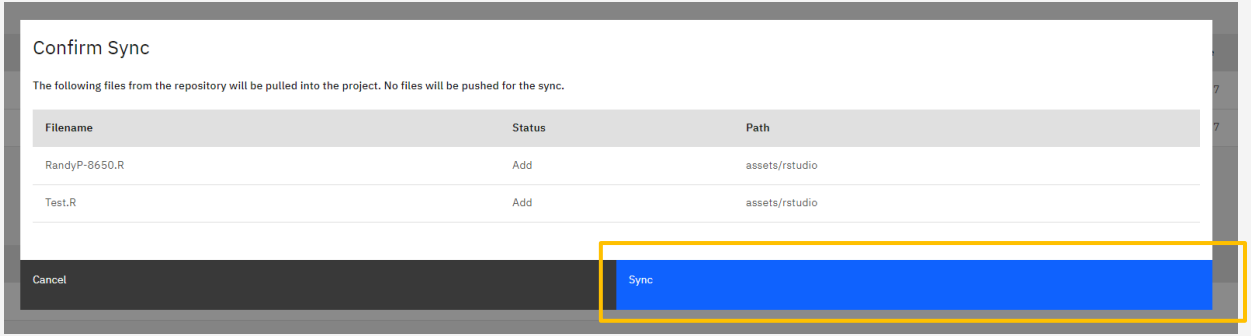
- Launch the JupyterLab IDE
- Pull from repo
- Add a Jupyter Notebook
- Push to repo

Lesson

- Understand how the different file environments (Watson Studio, RStudio and Git repo work together)

Watson Studio – Check Jupyter Notebook



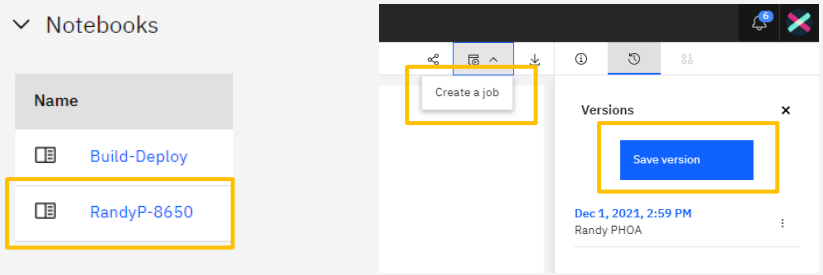


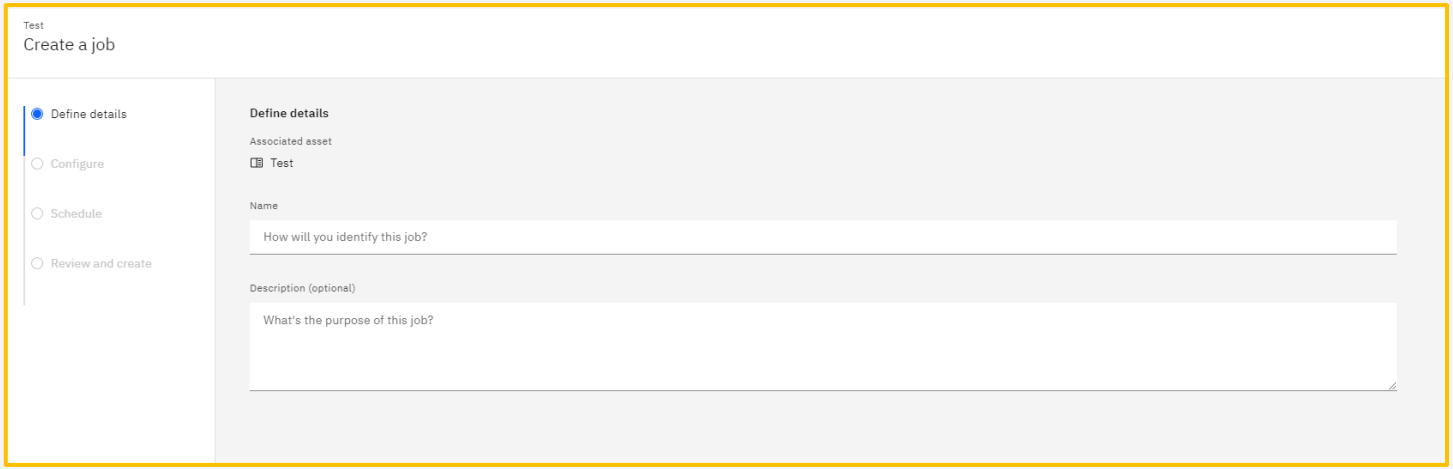
Notebooks

New Notebook +

Name	Shared	Scheduled	Status	Language	Last editor	Last modified
Build-Deploy				Python 3.7	Randy PHOA	Dec 02, 2021
RandyP-8650				Python 3.7	Randy PHOA	Dec 02, 2021

Run jobs in Watson Studio – Jupyter Notebook





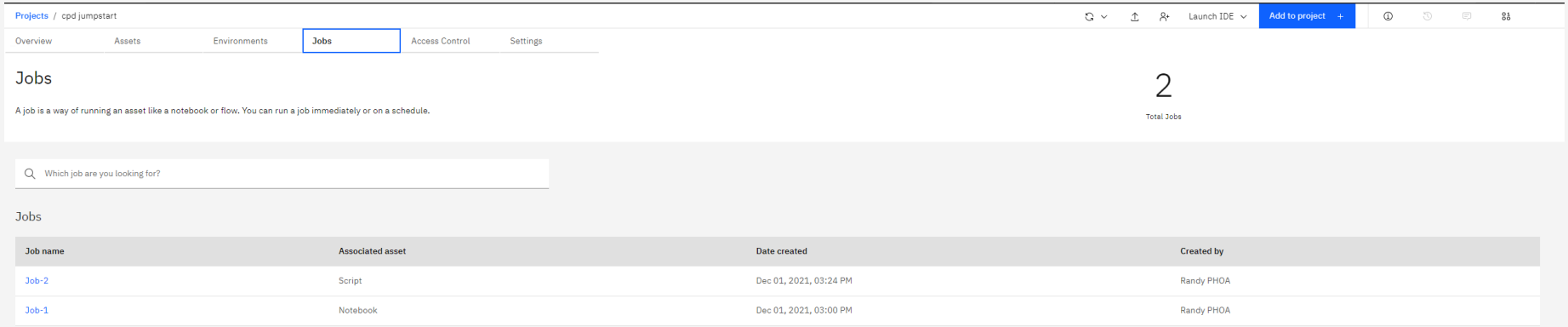
Exercise

- Sync project with repo
- Ensure you can see your Jupyter Notebook
- Create a Job

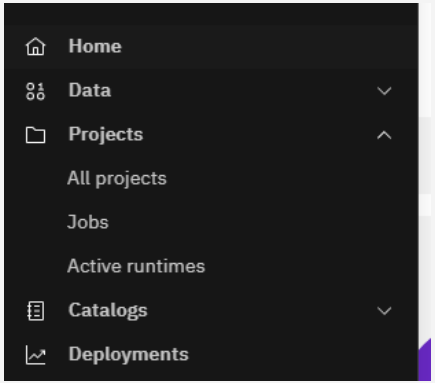
Lesson

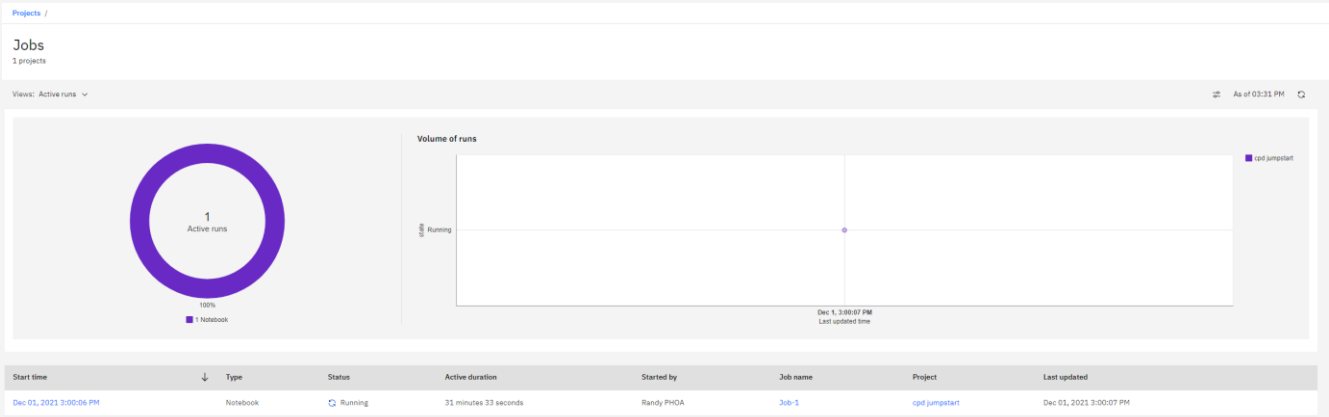
- Understand how the different file environments (Watson Studio, RStudio and Git repo work together)
- Learn how to create and configure a Job

Watson Studio – View job at project level



Watson Studio – View jobs at global level





Exercise

- View the jobs that you have created earlier

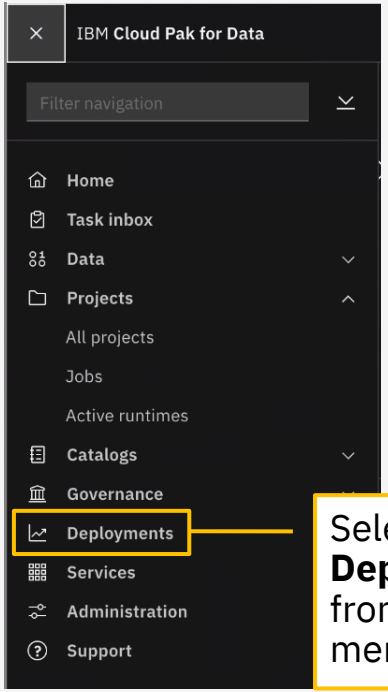
Lesson

- Learn the different ways of viewing job status

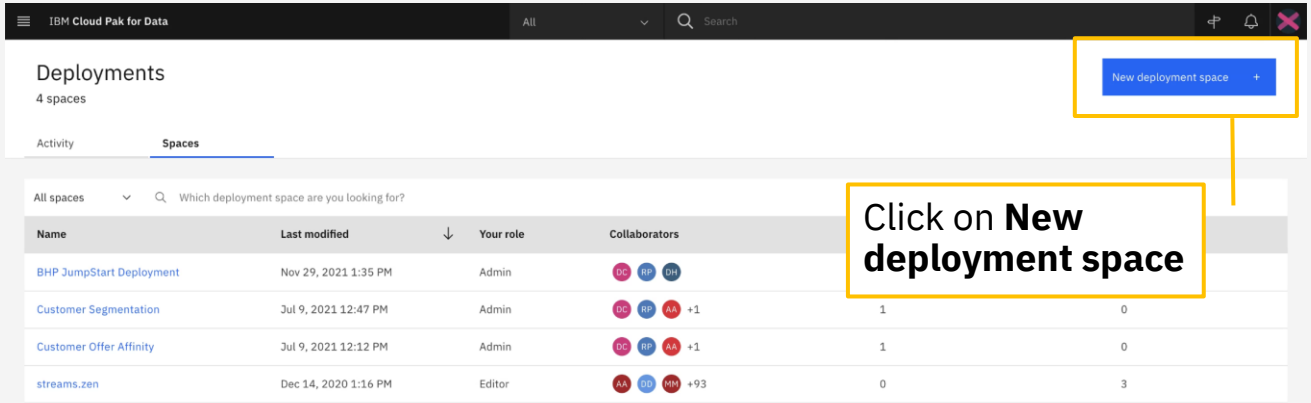
5. Deploy models

- Deployment spaces
- Promote assets
- Online scoring
- Batch scoring
- Schedule jobs

Create a deployment space

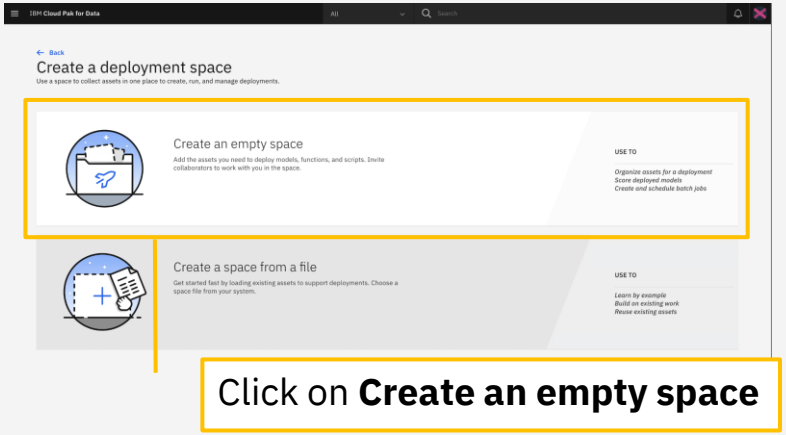


Select **Deployments** from the menu

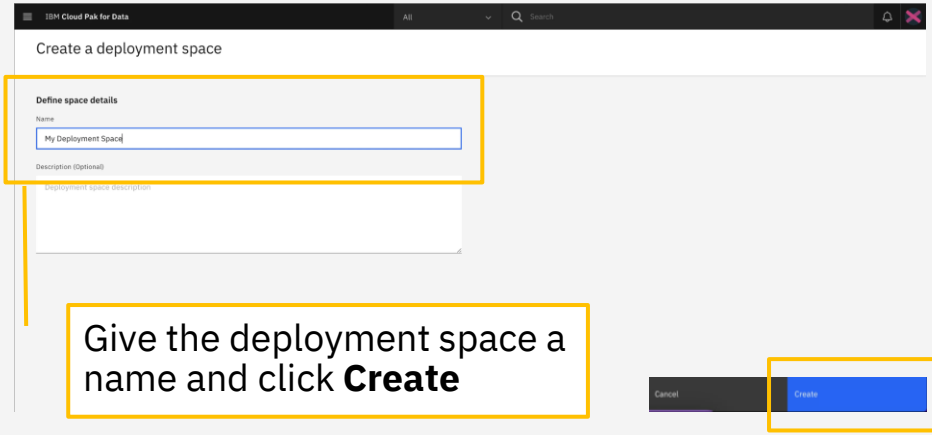


Click on **New deployment space**

Create a deployment space



Click on **Create an empty space**



Give the deployment space a name and click **Create**

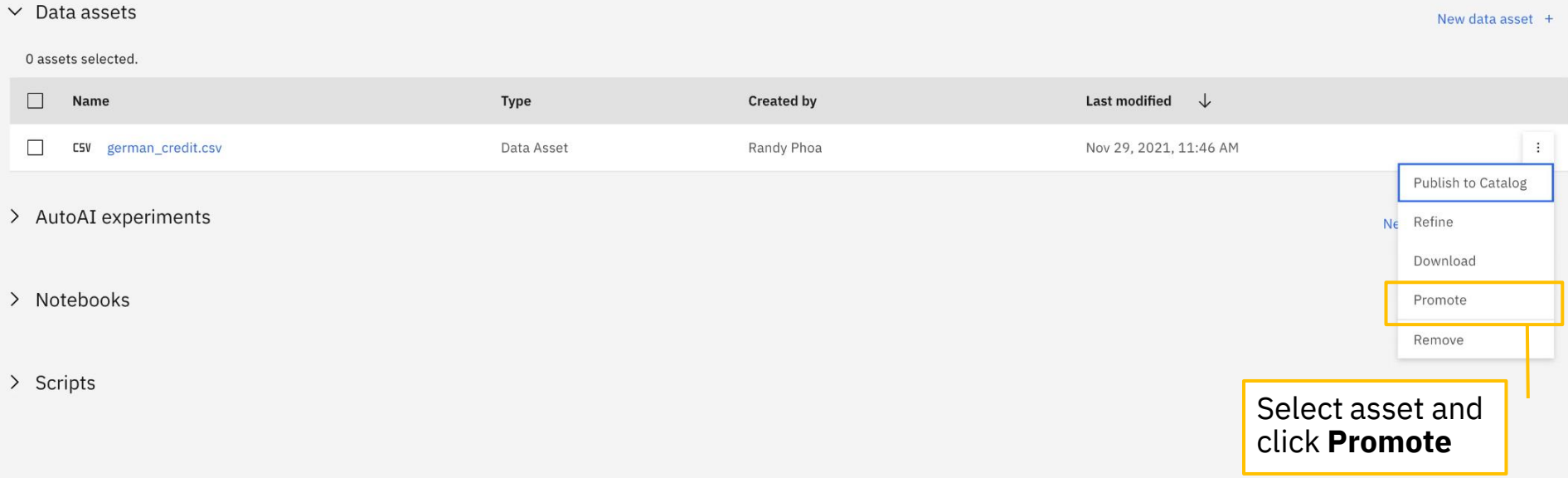
Exercise

- Create a deployment space

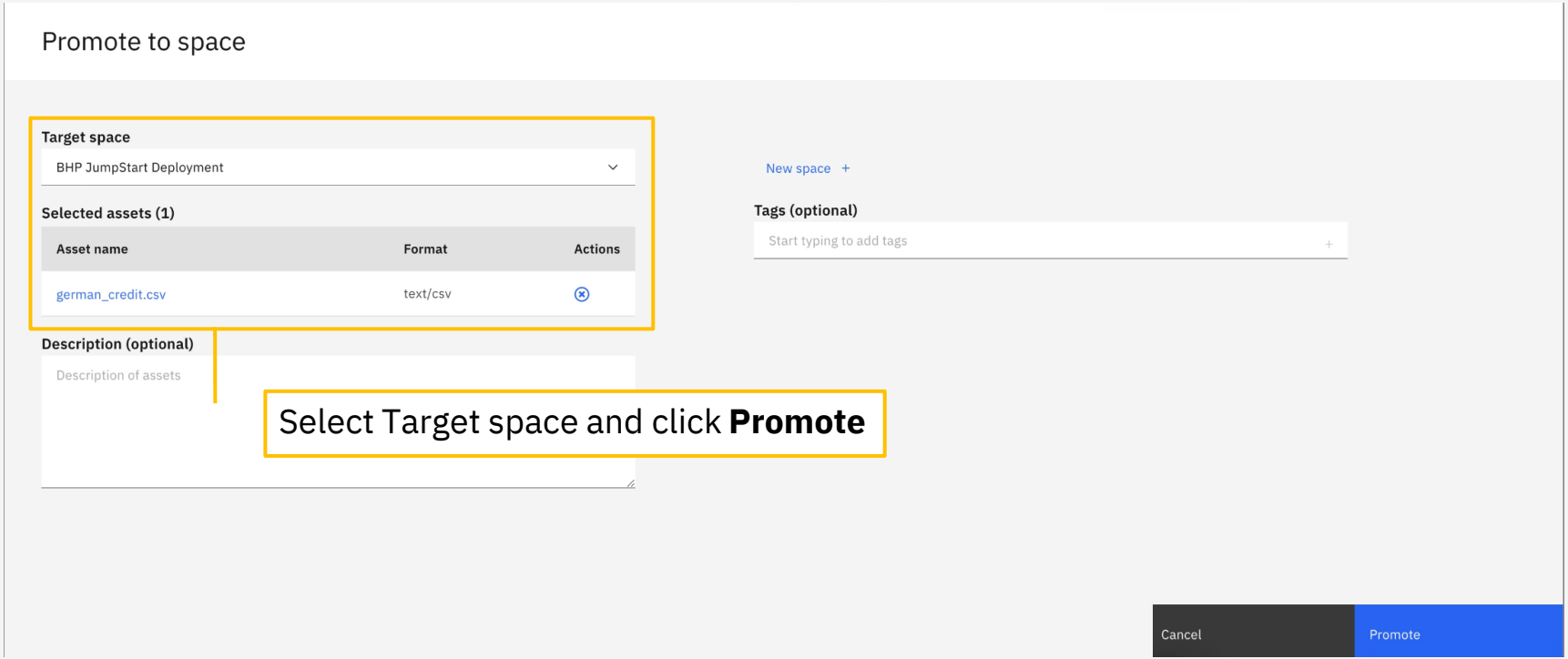
Lesson

- Understand what is a deployment space

Promote assets



Promote assets



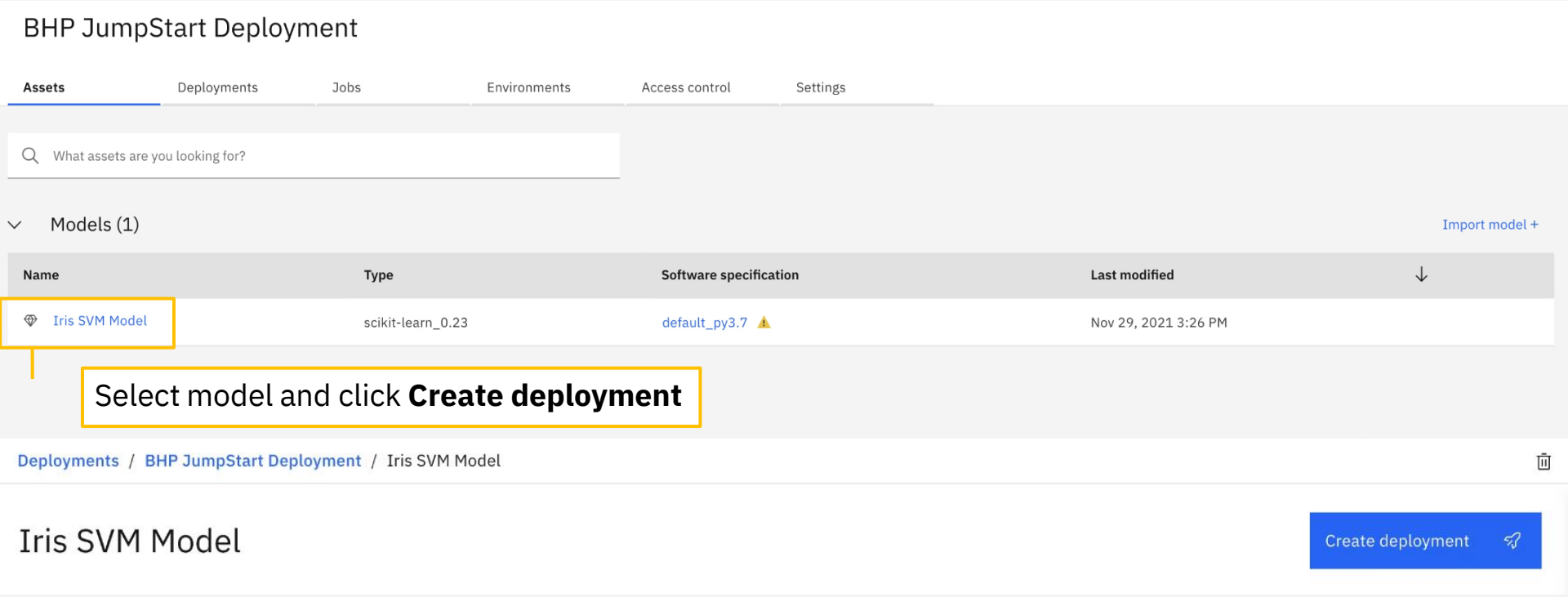
Exercise

- Promote your sample data file and your R or Python Script

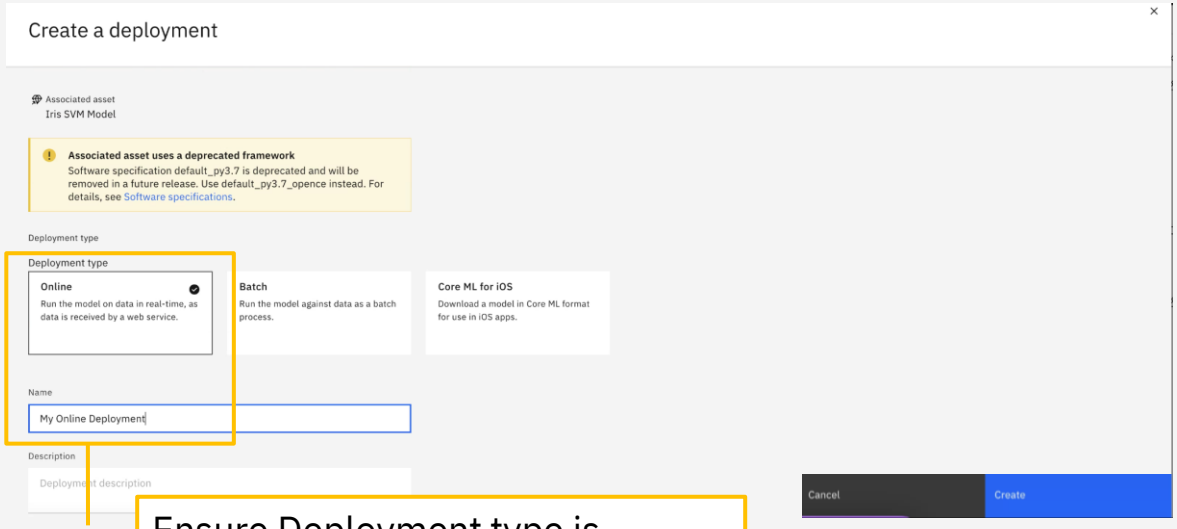
Lesson

- Understand how the deployment space (Watson Machine Learning) environment relates to Watson Studio, IDEs (RStudio and JupyterLab) and Git repo.

Online Scoring

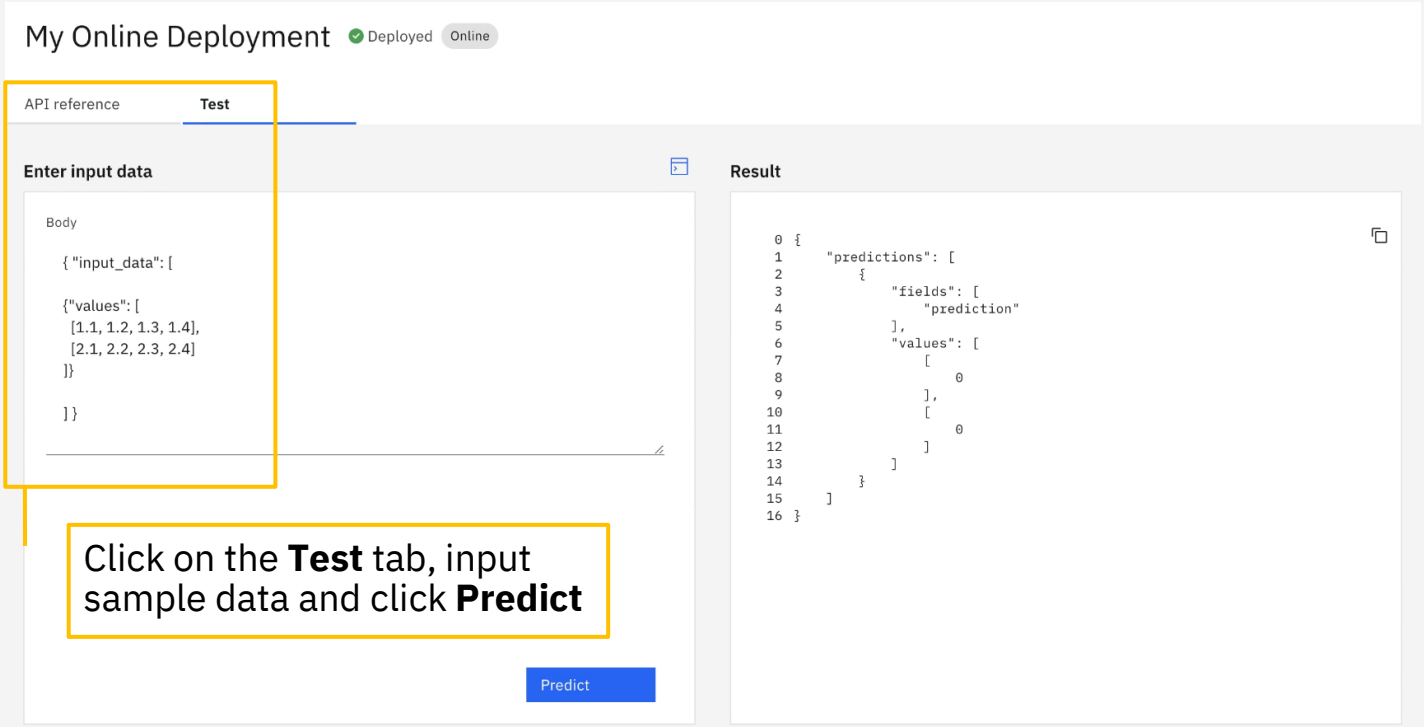


Online Scoring



Ensure Deployment type is **Online**, give your deployment a **name** and click **Create**

Online Scoring



Click on the **Test** tab, input sample data and click **Predict**

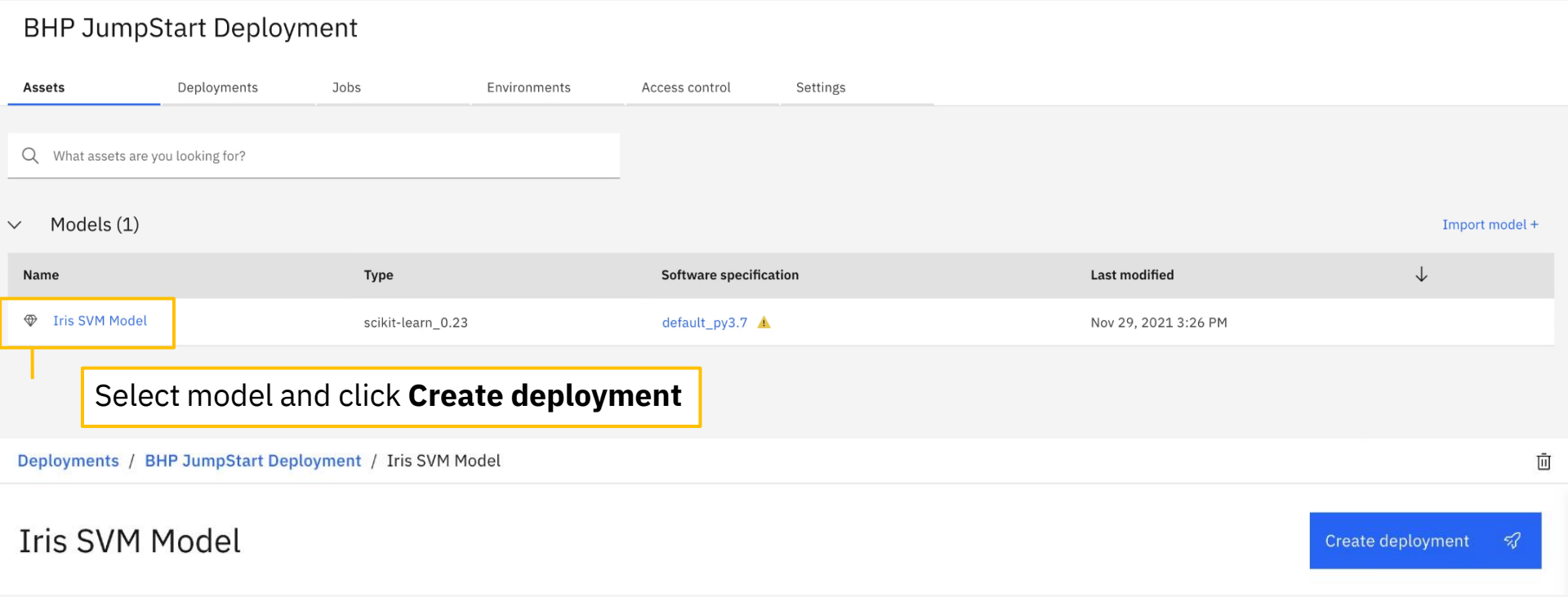
Exercise

- Create an online scoring deployment

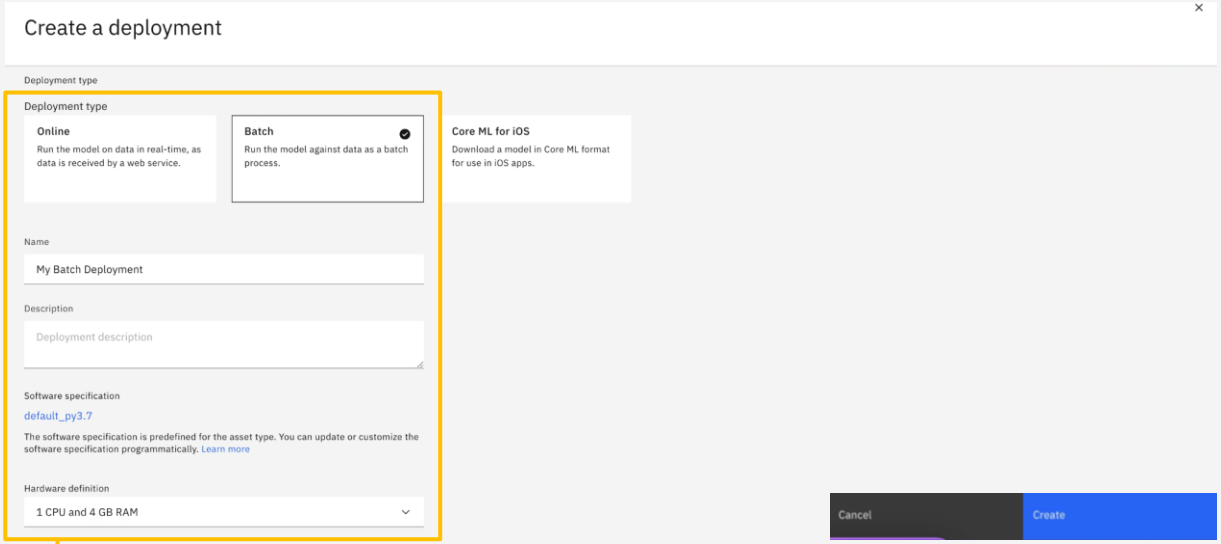
Lesson

- Learn how to configure an online scoring deployment and test using sample input data

Batch Scoring

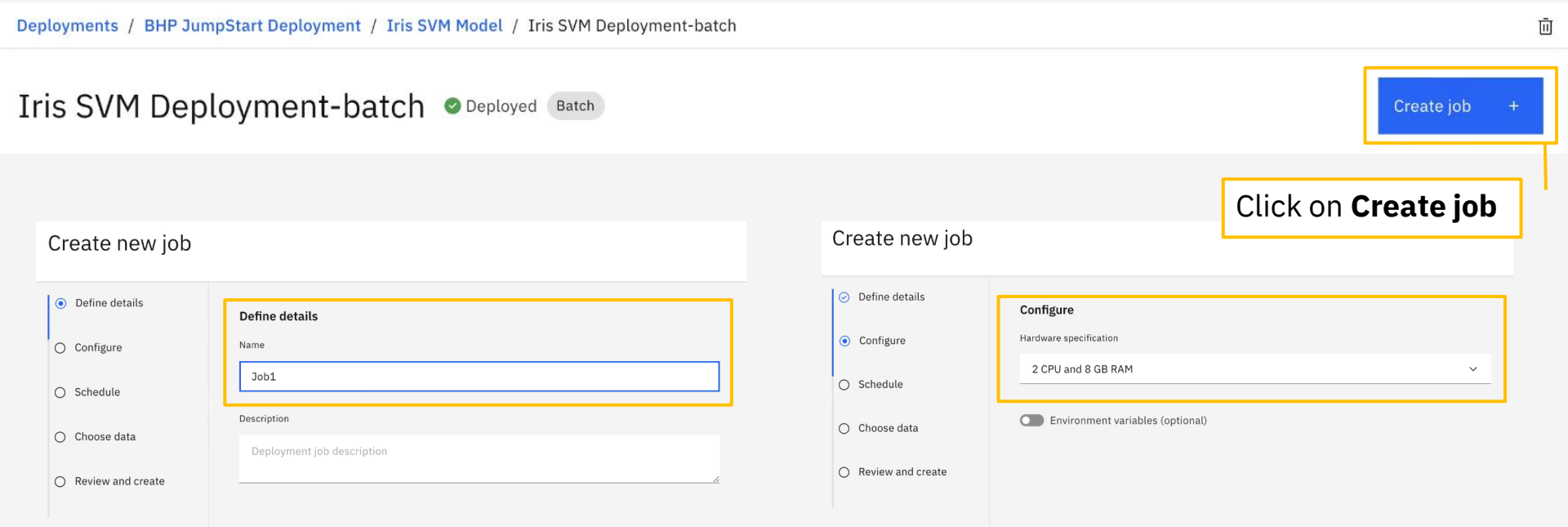


Batch Scoring

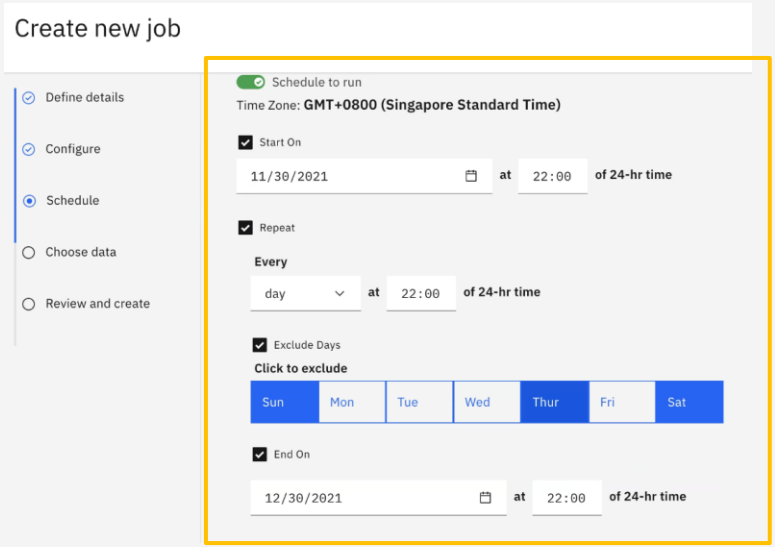


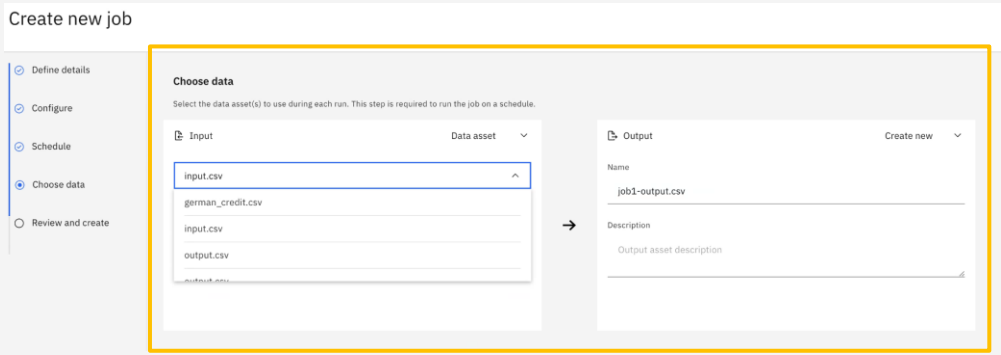
Ensure Deployment type is **Batch**, give your deployment a **name**, select the required **hardware definition** and click **Create**

Batch Scoring

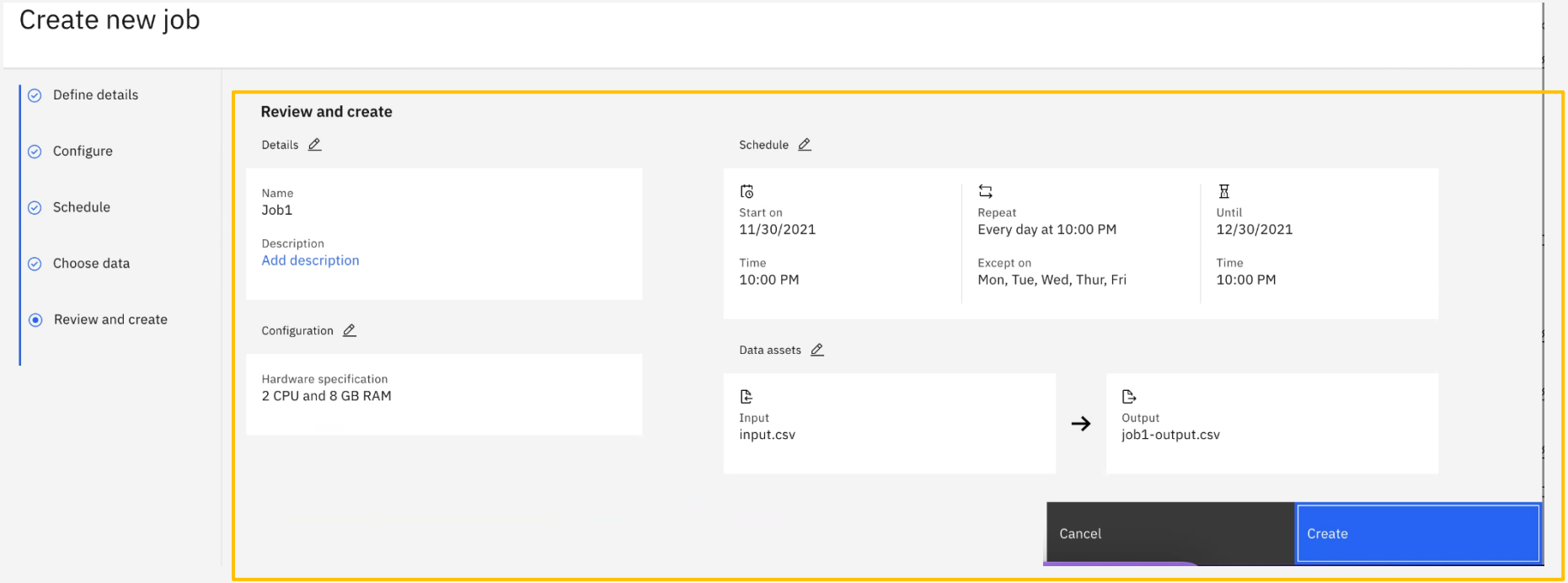


Batch Scoring

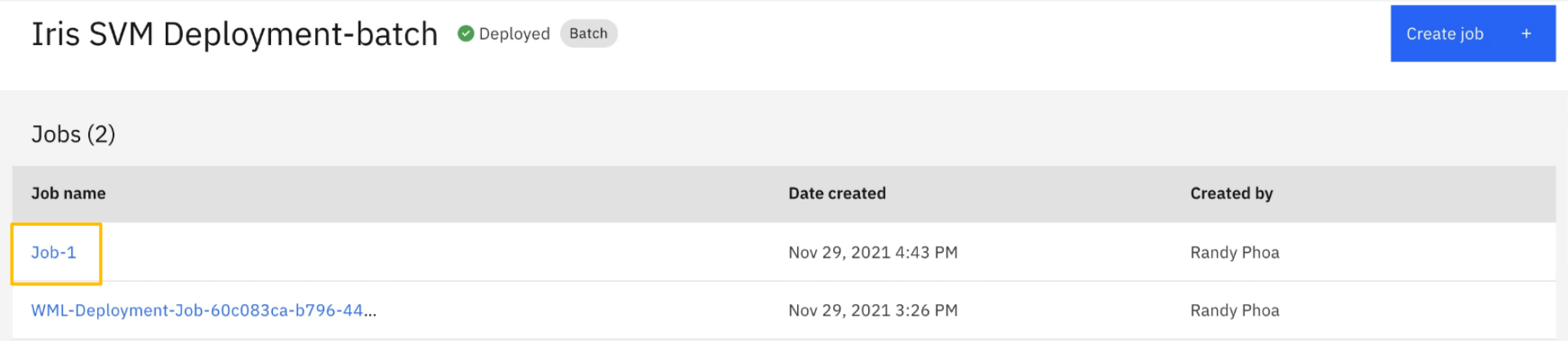




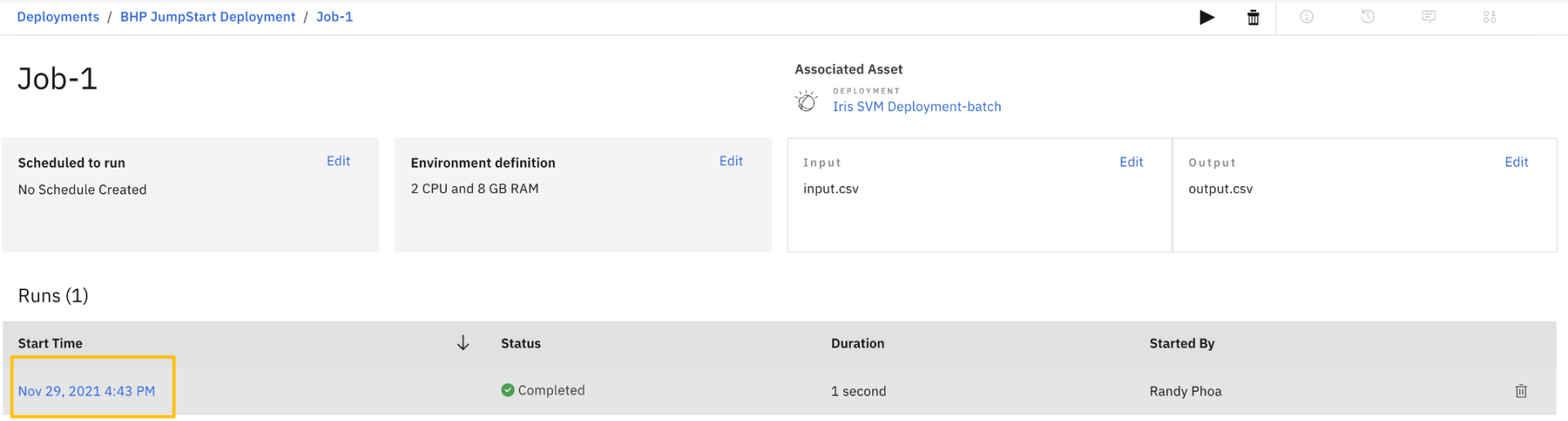
Batch Scoring



Batch Scoring – Check job status



Batch Scoring – Check job status



Batch Scoring – Check job status

Exercise

- Create a batch scoring deployment

Lesson

- Learn how to configure a batch scoring deployment and test using sample input data
- Learn how to check a batch job status

6. Resources

- Documentation – https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0
- Examples – https://github.com/IBM/watson-machine-learning-samples/tree/master/cpd3.5/notebooks/python_sdk