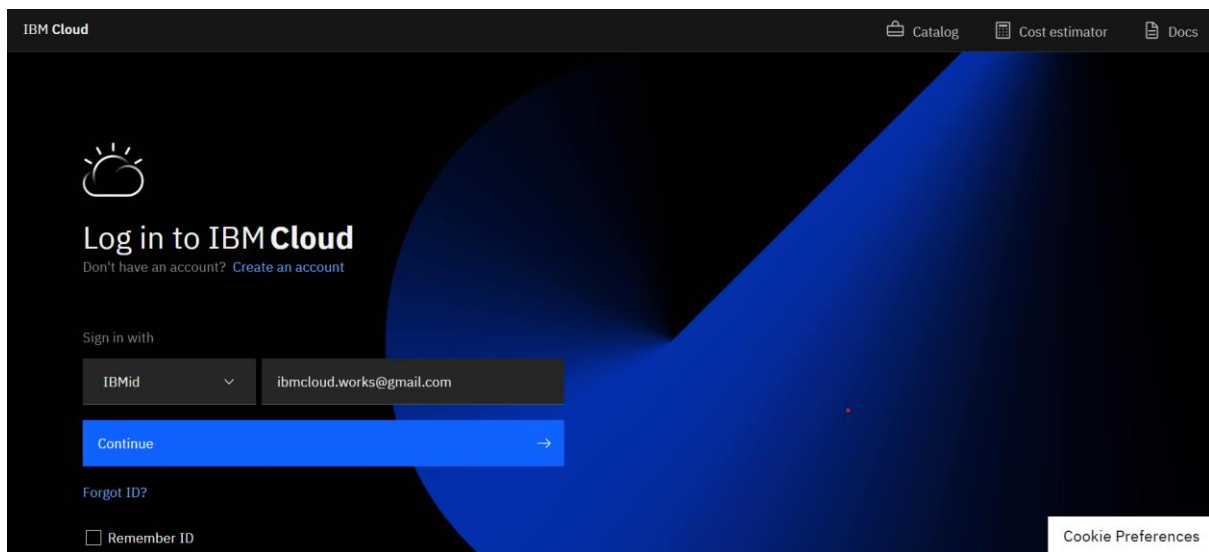
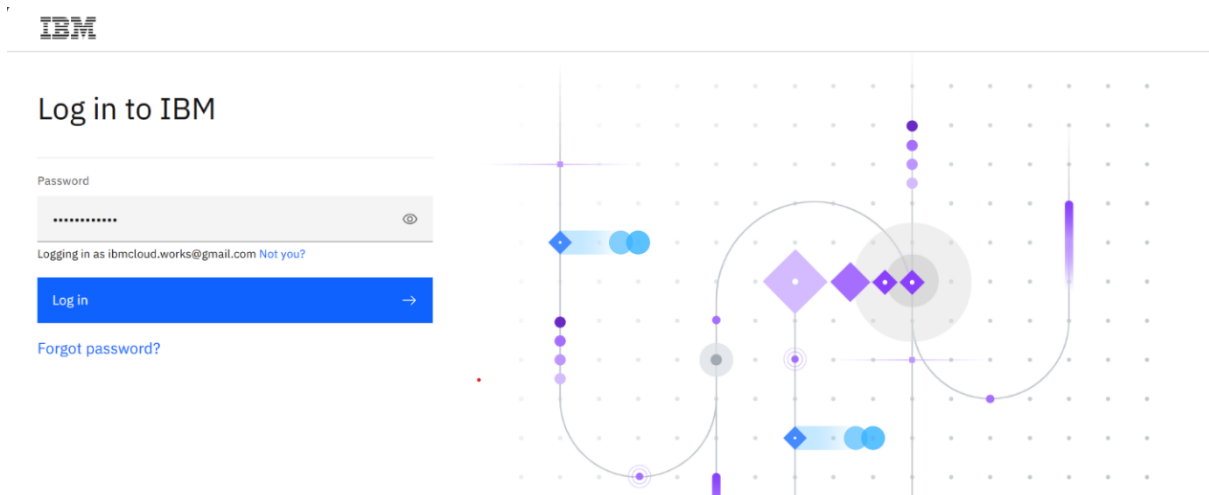


DATA PREP KIT

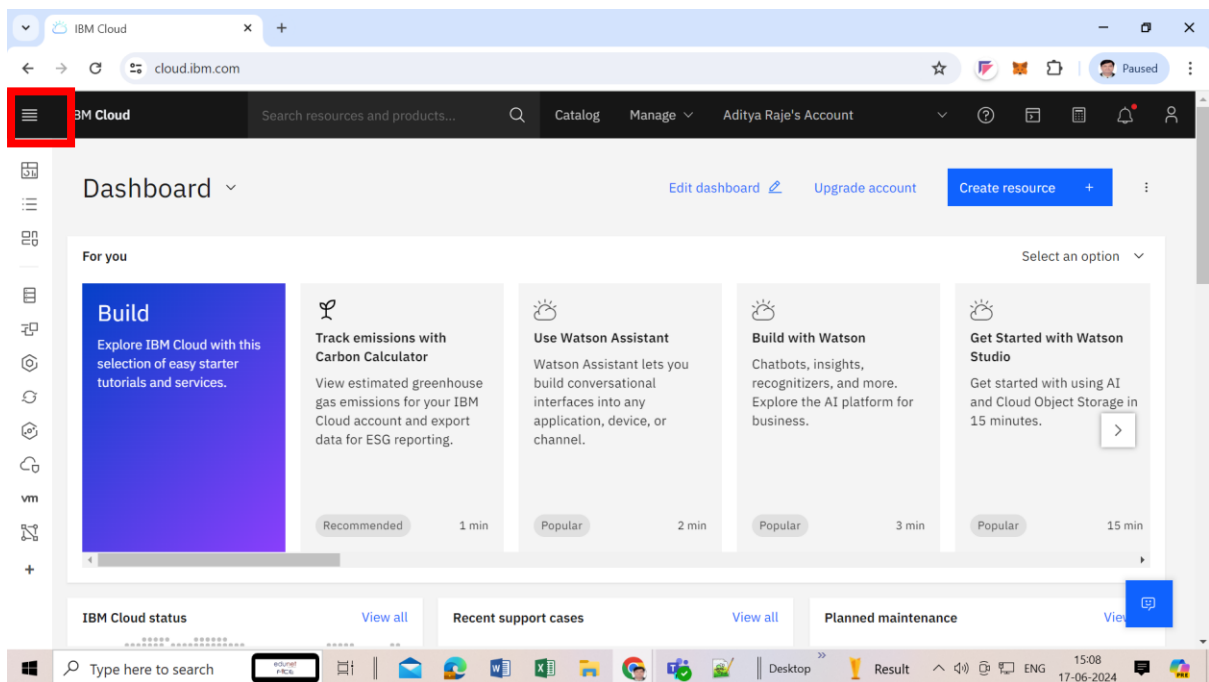
Step1: Open IBM Cloud login page with this link cloud.ibm.com, enter your Gmail and click on Continue



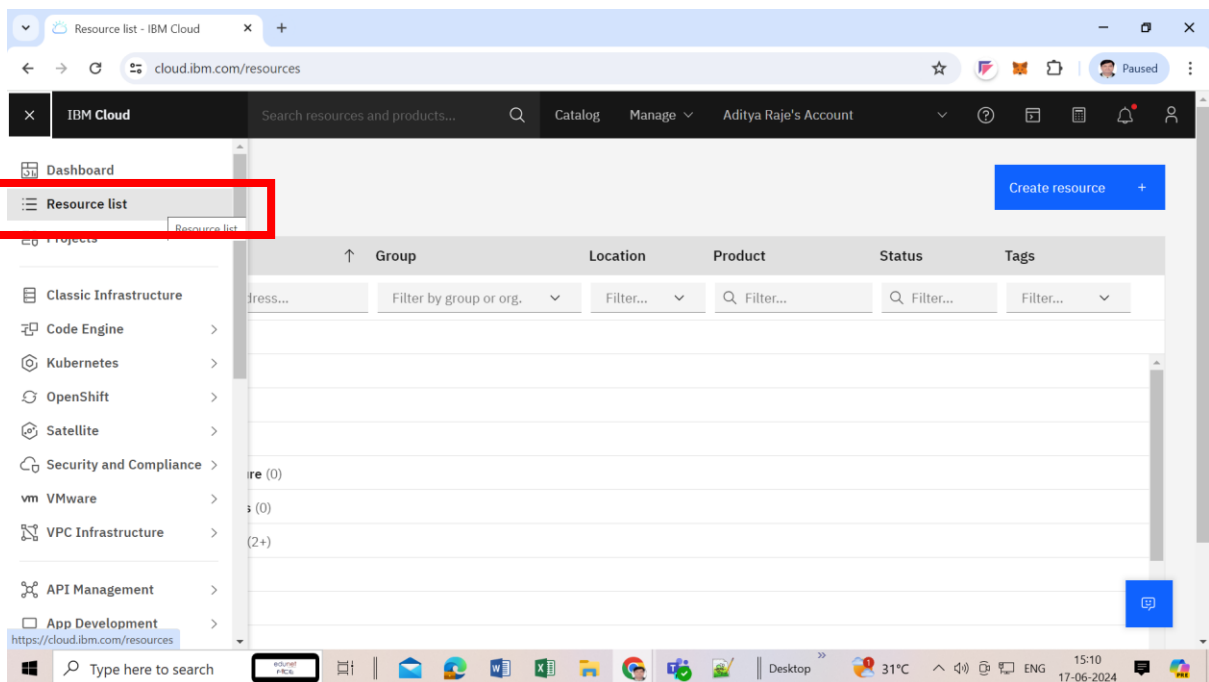
Step 2: Enter your IBM Academic portal password, Click on Login



Step 3: This is IBM Cloud Dash board



Step4 : From top left ,Click on Navigation Menu → Resource list



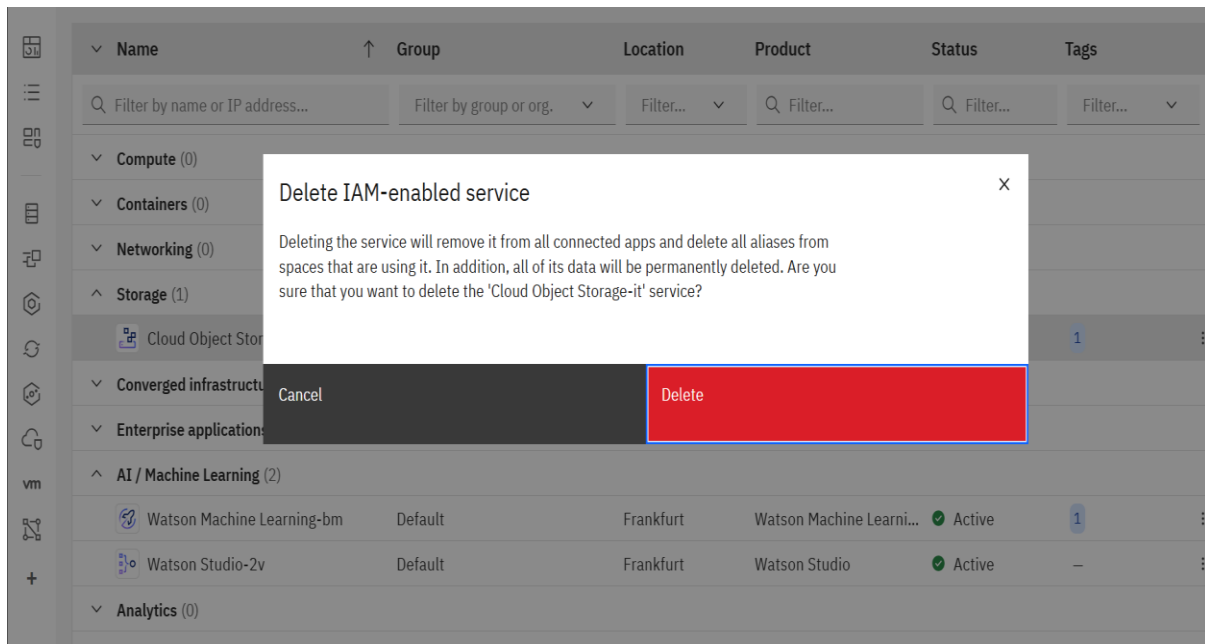
Step 5: Make sure you deleted all resources under **Storage** and **AI/ Machine Learning**

The screenshot shows the IBM Cloud 'Resource list' page. The browser address bar displays 'cloud.ibm.com/resources'. The page header includes the IBM Cloud logo, a search bar, and navigation links for 'Catalog', 'Manage', and the user's account 'Aditya Raje's Account'. On the left, a sidebar contains icons for various resource categories. The main content area features a table with columns: Name, Group, Location, Product, Status, and Tags. Above the table are filter inputs for each column. The table is currently empty, with counts in parentheses next to the category names in the left sidebar: Compute (0), Containers (0), Networking (0), Storage (0), Converged infrastructure (0), Enterprise applications (0), AI / Machine Learning (0), Analytics (0), Blockchain (0), and Databases (0). A blue 'Create resource' button is located in the top right corner.

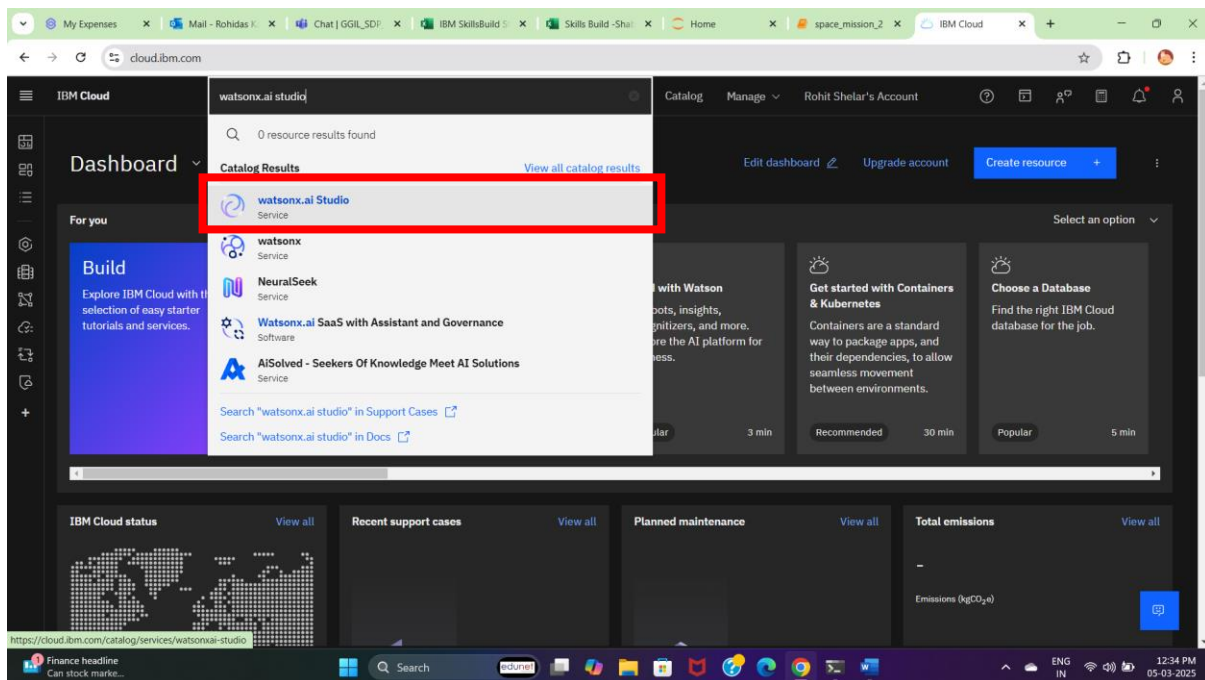
This screenshot shows the IBM Cloud 'Resource list' page with resources populated. The 'Storage' category now shows 1 resource, and 'AI / Machine Learning' shows 2 resources. The table lists the following resources:

Name	Group	Location	Product	Status	Tags
Cloud Object Storage-it	Default	Global	Cloud Object Storage	Active	1
Watson Machine Learning-bm	Default	Frankfurt	Watson Machine Learn...	Active	1
Watson Studio-2v	Default	Frankfurt	Watson Studio	Active	-

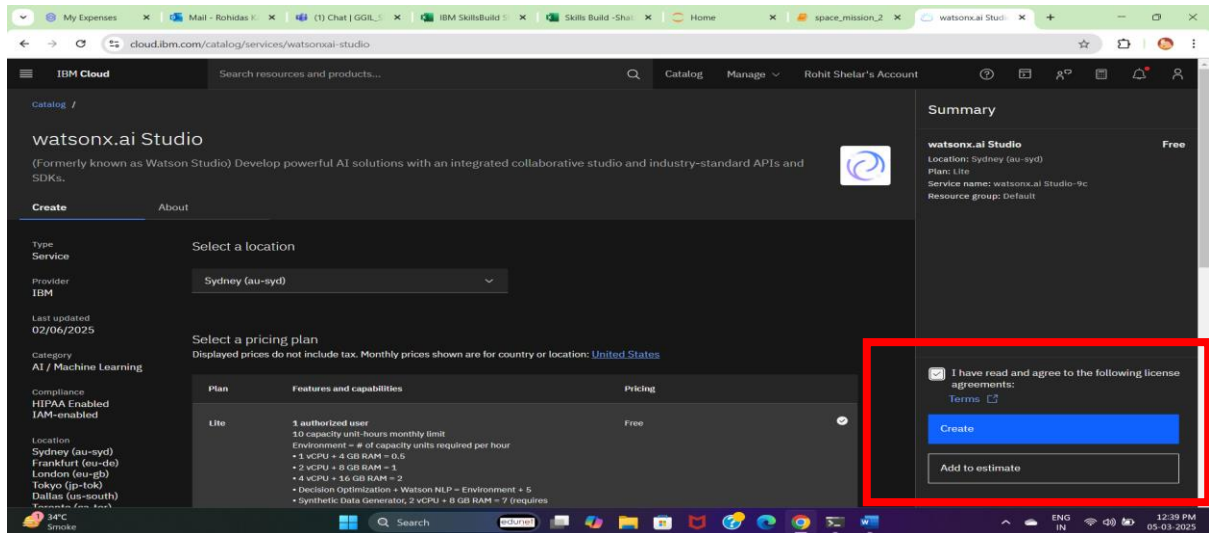
This screenshot shows the IBM Cloud 'Resource list' page with the context menu open for the 'Cloud Object Storage-it' resource. The menu options are: Edit name, Edit tags, Export access report, and Delete. The 'Delete' option is highlighted in red.



Step 6: Search for Watsonx.ai Studio service



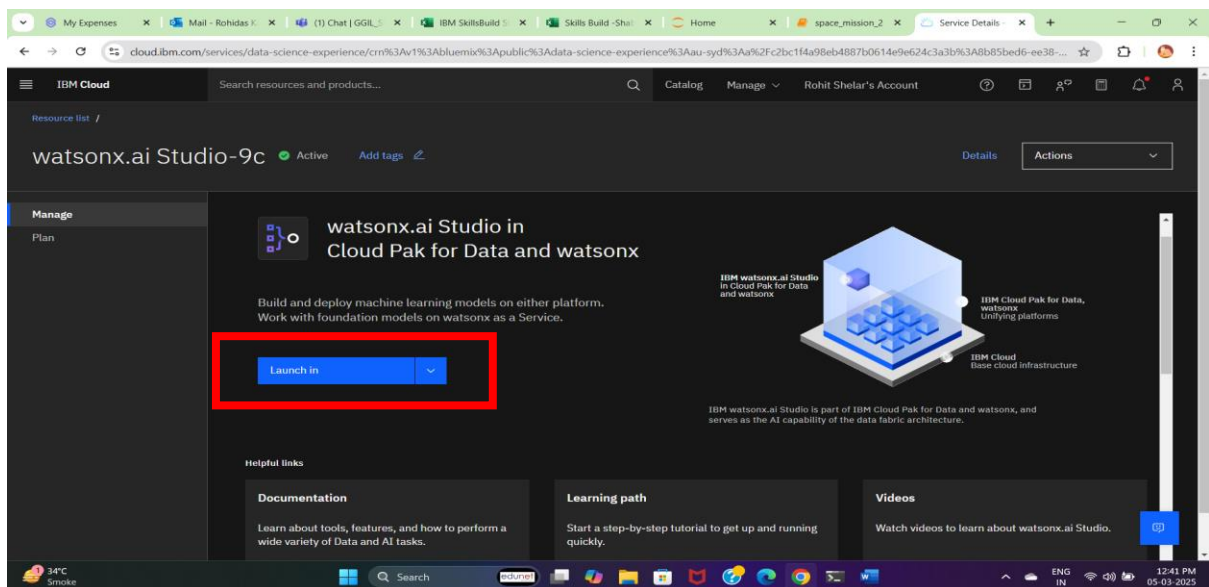
Step 7: Create Watsonx.ai Studio service with Free pricing, Click on check box then click on create



The screenshot shows the IBM Cloud catalog page for Watsonx.ai Studio. The 'Create' tab is active, and the 'Select a location' dropdown is set to 'Sydney (au-syd)'. The 'Select a pricing plan' dropdown is set to 'Free'. A red box highlights the 'I have read and agree to the following license agreements' checkbox, which is checked, and the 'Create' button below it.

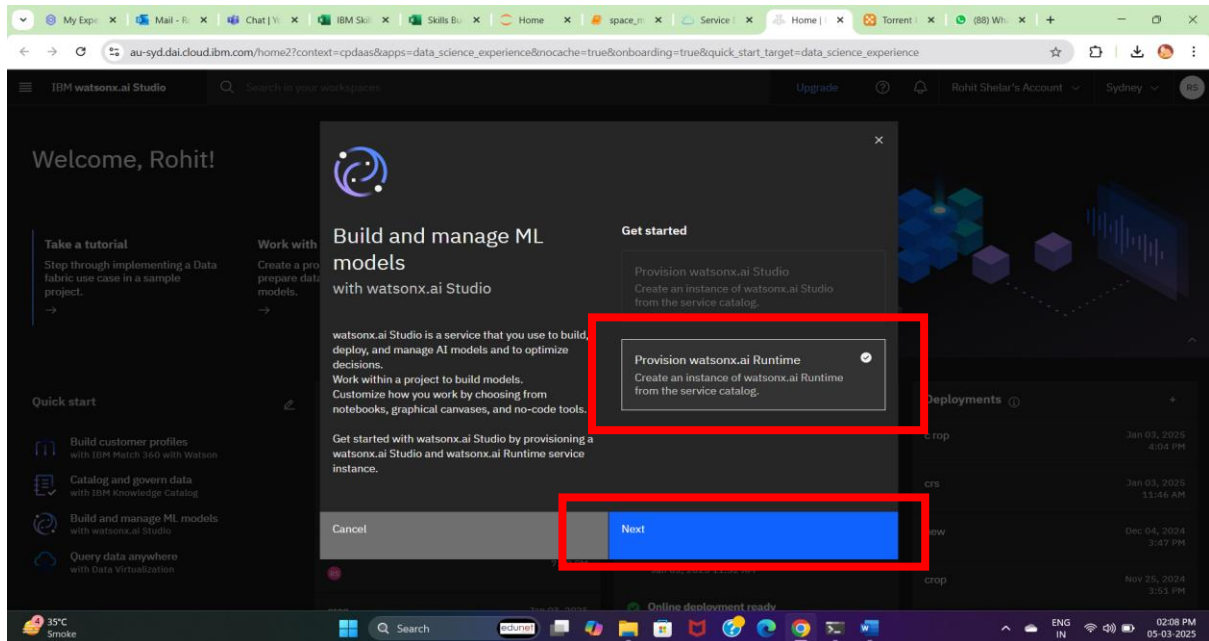
Plan	Features and capabilities	Pricing
Lite	1 authorized user 10 capacity unit-hours monthly limit Environment = # of capacity units required per hour • 1 vCPU + 4 GB RAM = 0.5 • 2 vCPU + 8 GB RAM = 1 • 4 vCPU + 16 GB RAM = 2 • Decision Optimization + Watson NLP = Environment + 5 • Synthetic Data Generator, 2 vCPU + 8 GB RAM = 7 (requires)	Free

Step 8: Click on Launch In

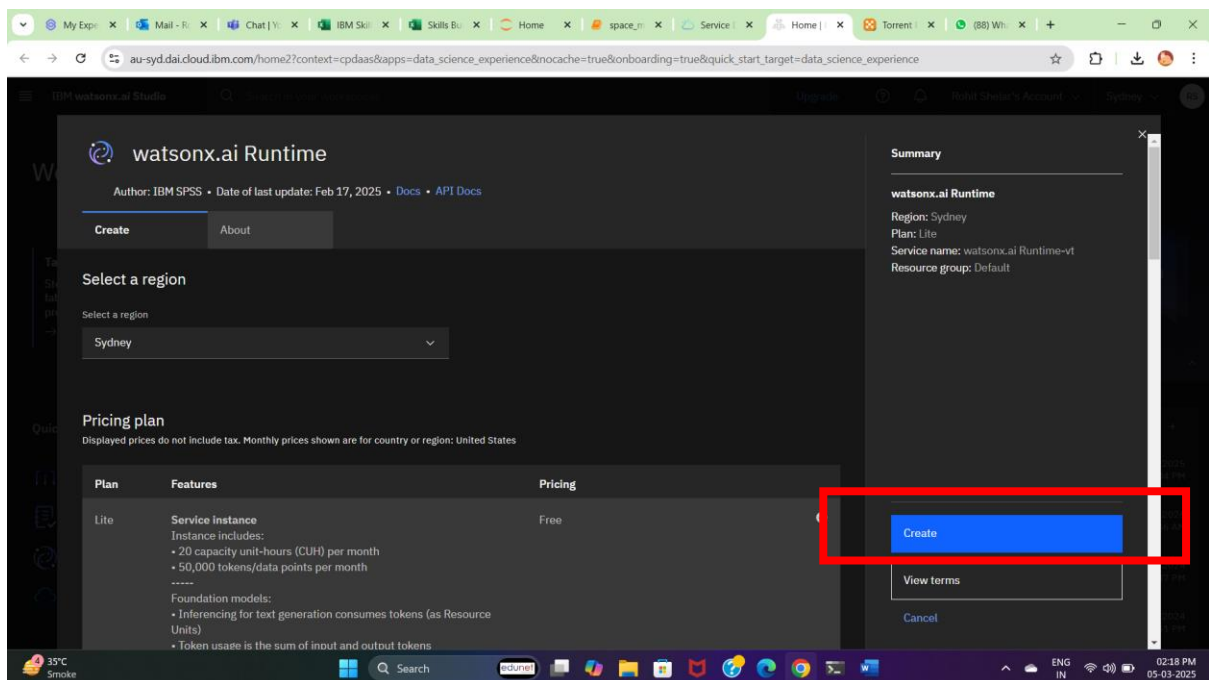


The screenshot shows the IBM Cloud 'Service Details' page for Watsonx.ai Studio-9c. The 'Launch in' button is highlighted with a red box. The page also includes a 'Manage' tab, a 'Plan' section, and a 'Helpful links' section with 'Documentation', 'Learning path', and 'Videos'.

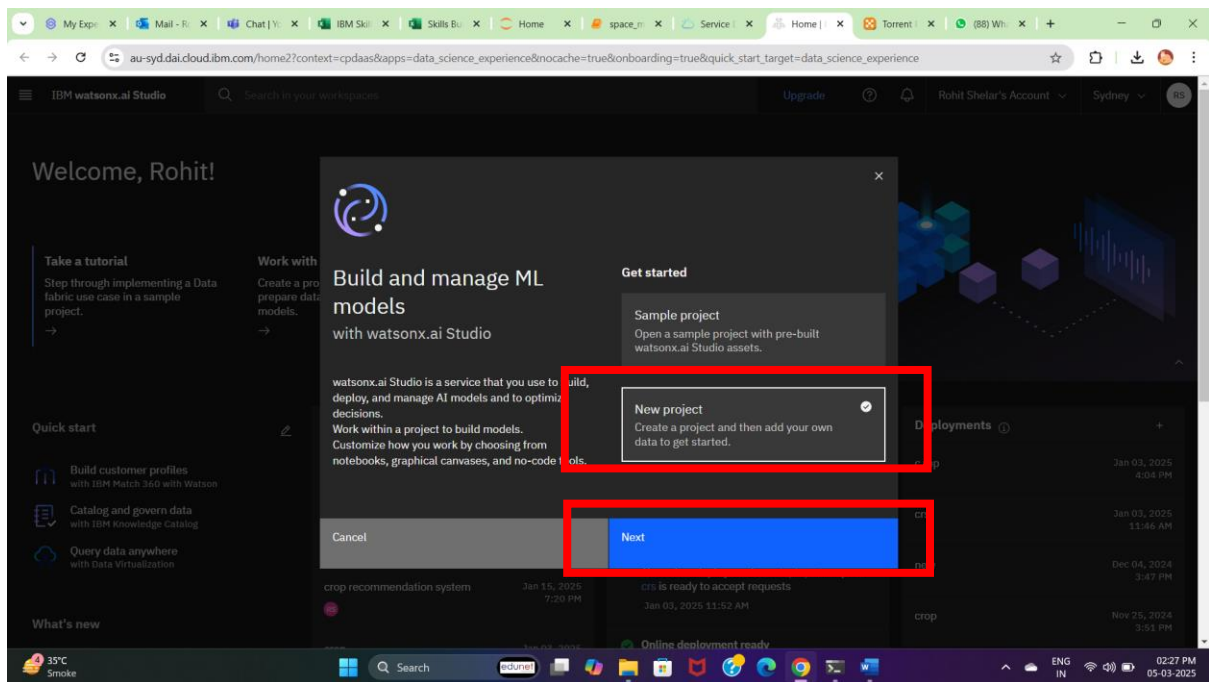
Step 8: Select Provision Watsonx.ai Runtime, Click on Next



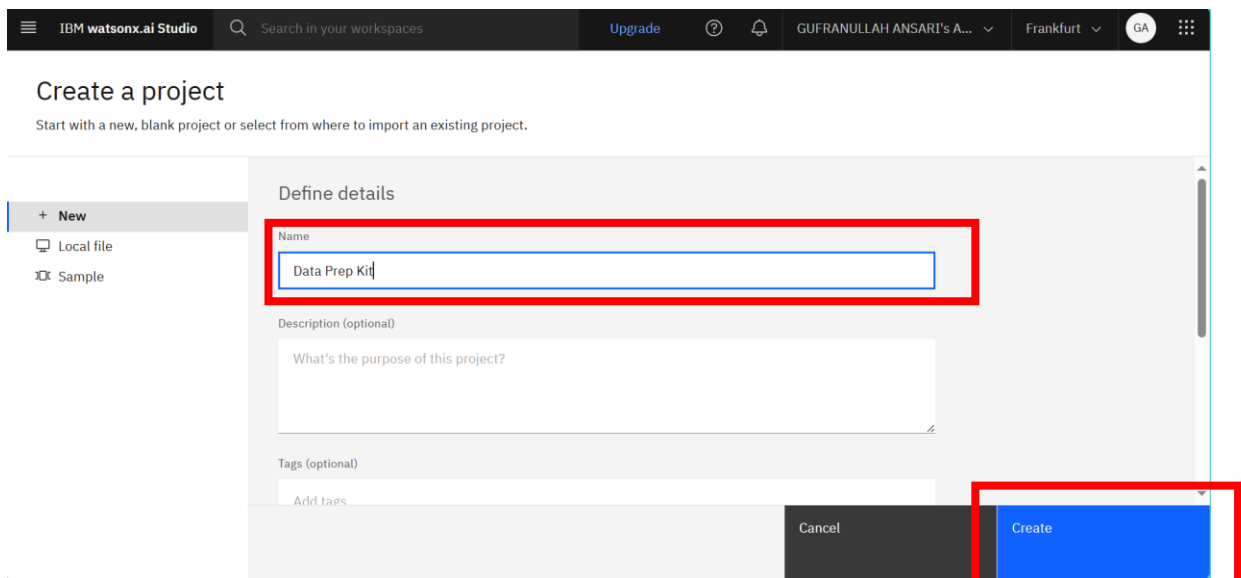
Step 9: Create Watsonx.ai Runtime service, click on Create



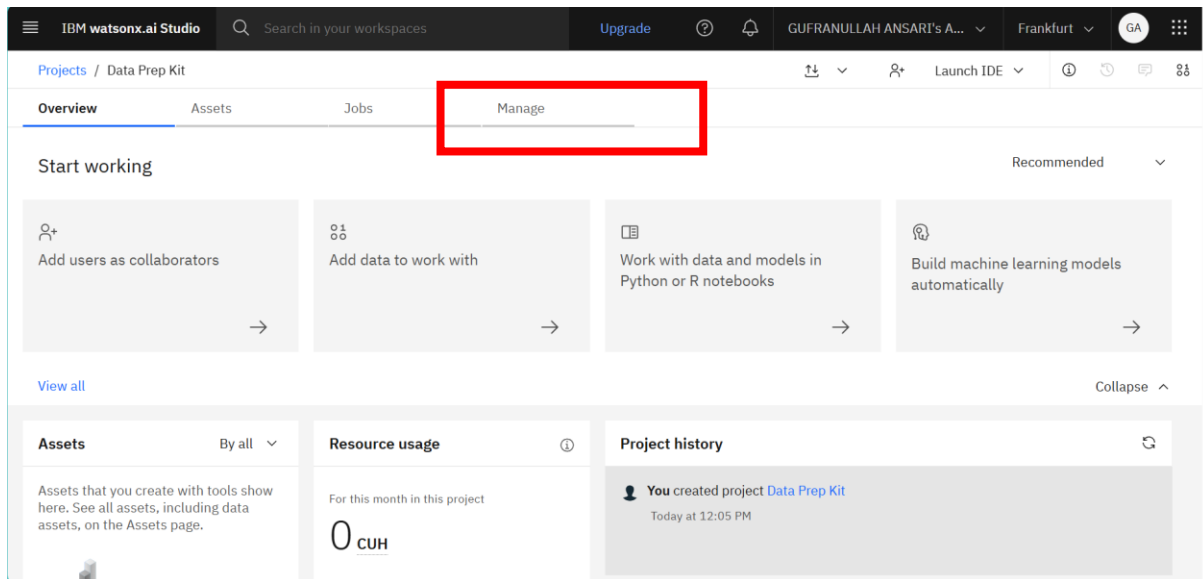
Step 10: Click on New Project and Next



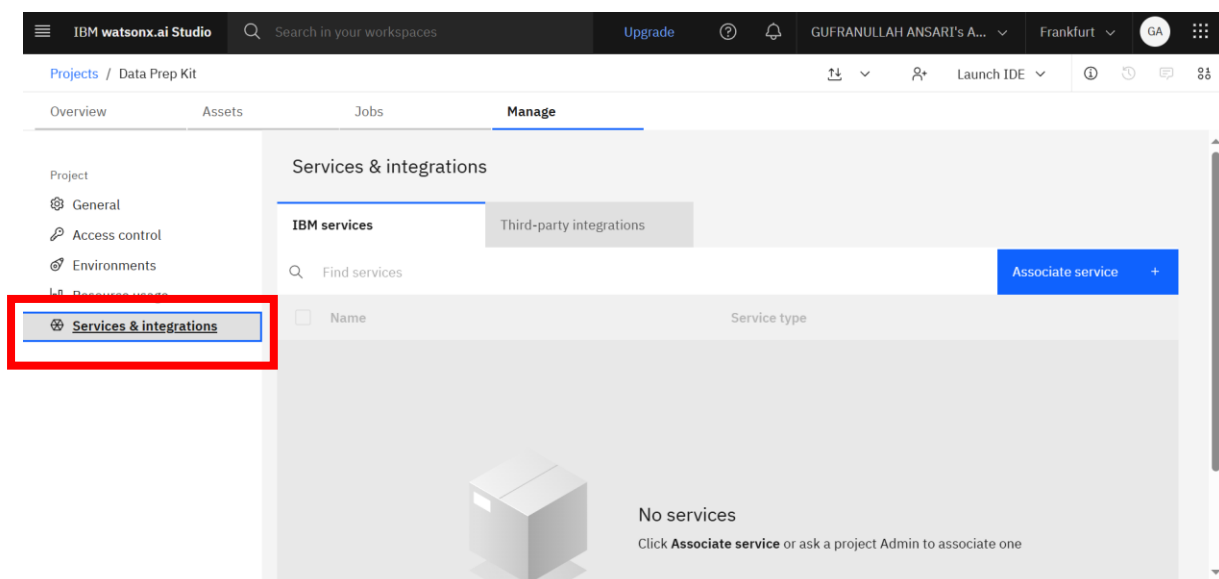
Step 11: In Create Project window – Provide details about Name, Description and click on Add for storage



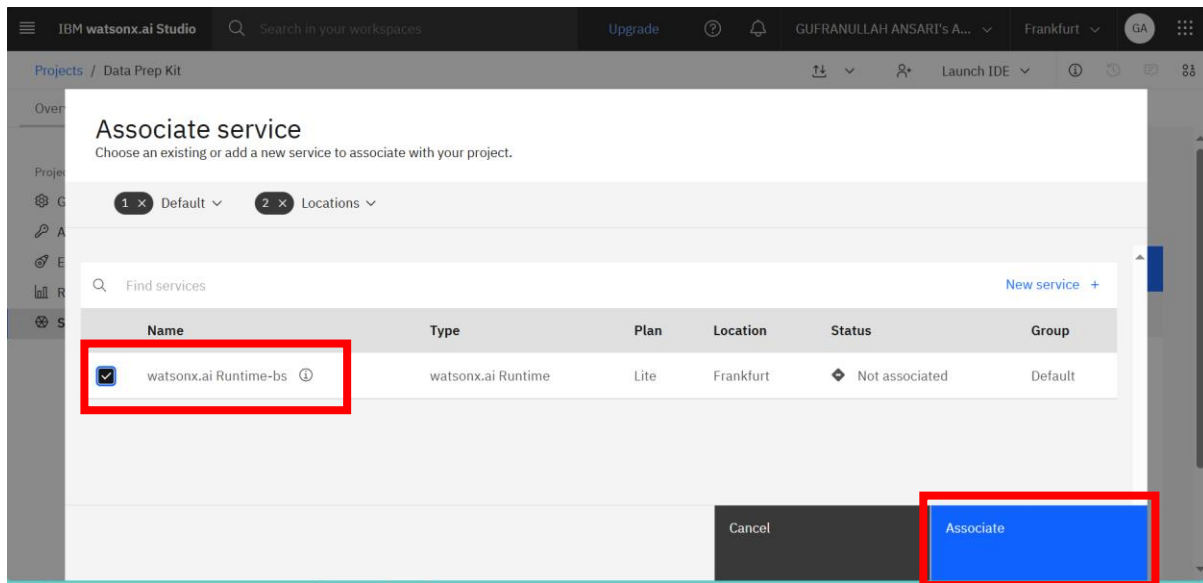
Step 12: Click on the Manage Section



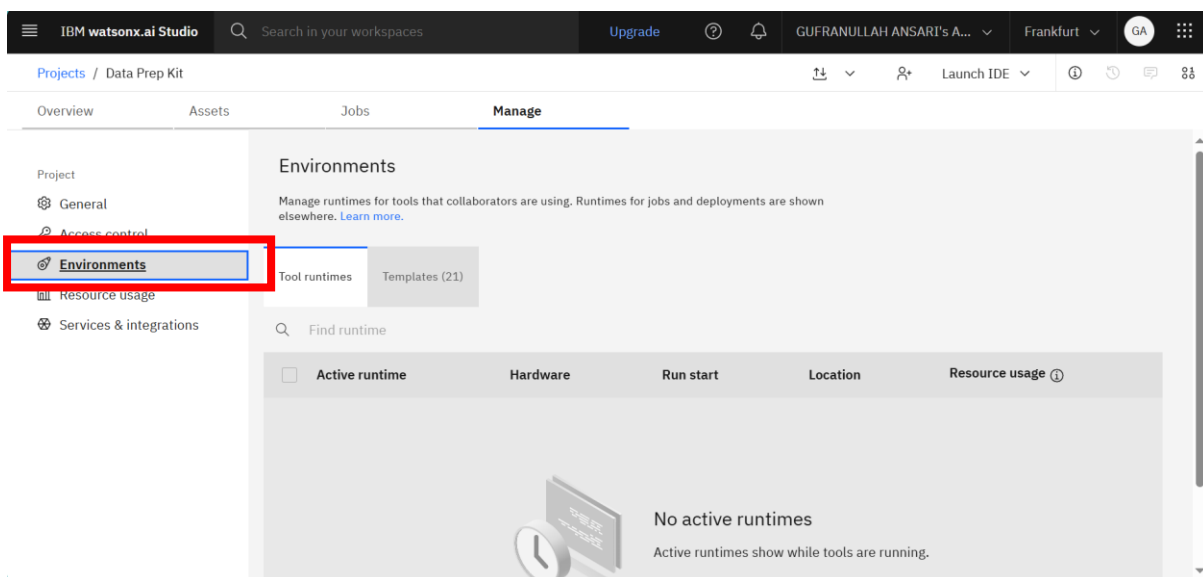
Step 13: Click on the Service & Integrations



Step 14: Associate the service watsonx.ai Runtime



Step 15: Select the Environments section



Step 16: Click on the Templates section

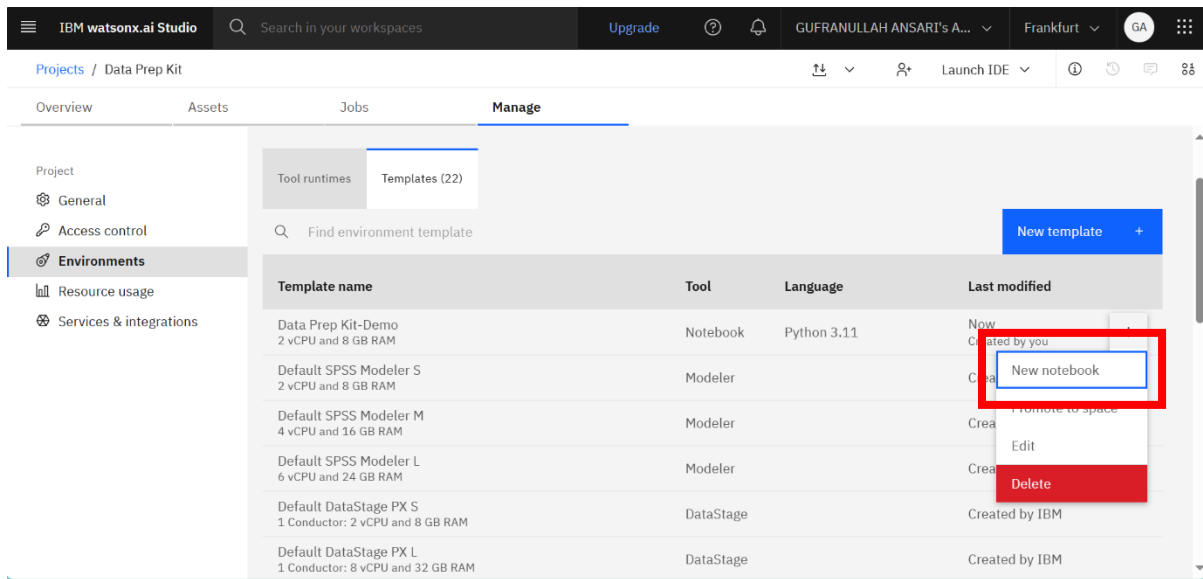
The screenshot shows the IBM watsonx.ai Studio interface. The top navigation bar includes the logo, a search bar, and user information. The main sidebar on the left lists navigation options: Overview, Assets, Jobs, and Manage (selected). Under 'Manage', there are sub-tabs: Overview, Assets, Jobs, and Manage. The 'Environments' section is active, displaying a list of tool runtimes. A red box highlights the 'Templates (21)' link in the 'Tool runtimes' section. Below this, a table lists various templates with columns for Template name, Tool, Language, and Last modified.

Template name	Tool	Language	Last modified
Default SPSS Modeler S 2 vCPU and 8 GB RAM	Modeler		Created by IBM
Default SPSS Modeler M 4 vCPU and 16 GB RAM	Modeler		Created by IBM
Default SPSS Modeler L 6 vCPU and 24 GB RAM	Modeler		Created by IBM
Default DataStage PX S 1 Conductor: 2 vCPU and 8 GB RAM	DataStage		Created by IBM

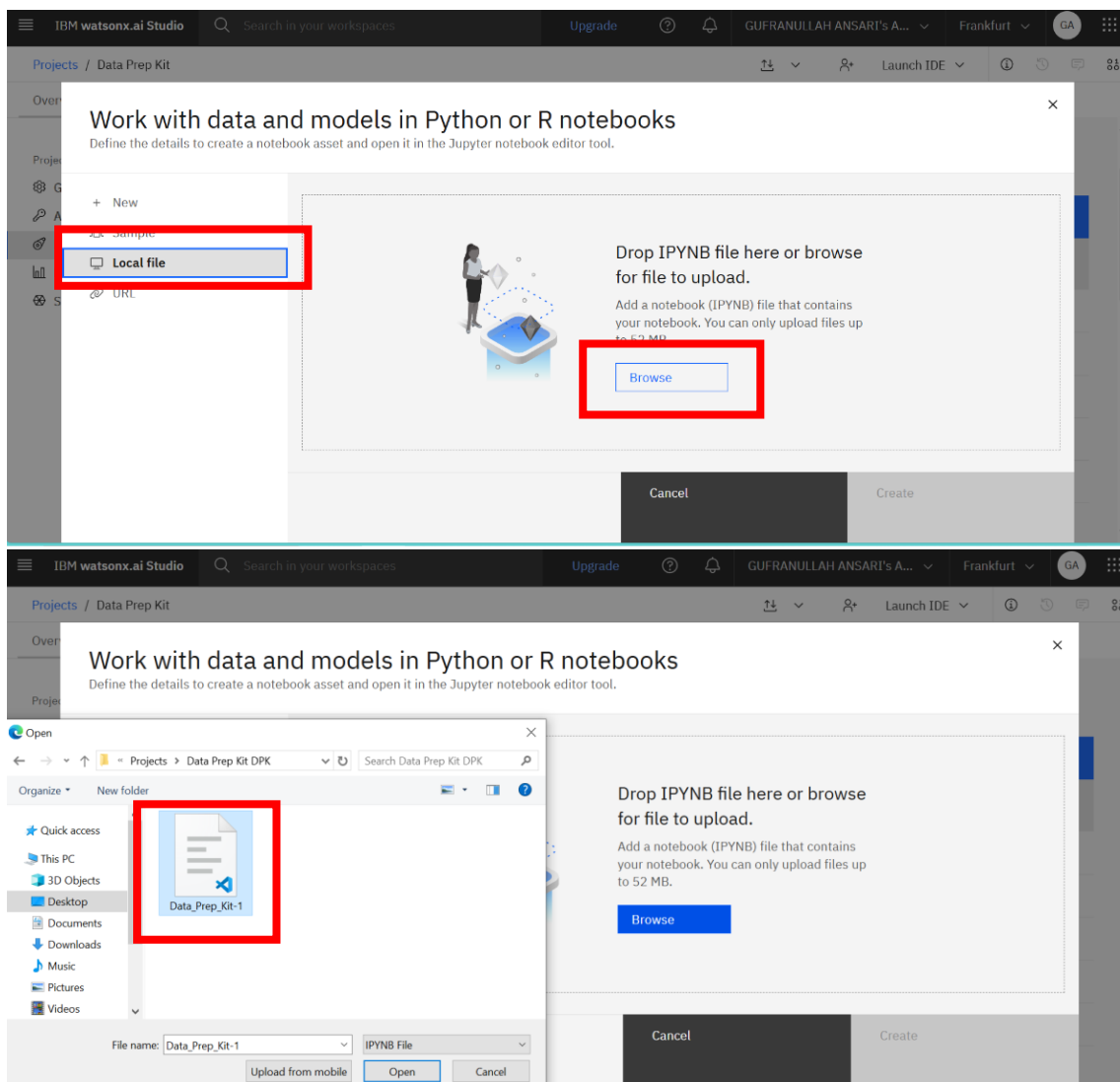
Step 17: Define New environment Name & Hardware Configuration Create

The screenshot shows the 'New environment' dialog box in IBM watsonx.ai Studio. The dialog is divided into two main sections: 'Define environment details' and 'Define configuration'. In the 'Define environment details' section, the 'Name' field is highlighted with a red box and contains the text 'Data Prep Kit-Demo'. In the 'Define configuration' section, the 'Hardware configuration' dropdown is highlighted with a red box and shows '2 vCPU and 8 GB RAM'. The 'Software version' dropdown shows 'Runtime 24.1 on Python 3.11'. At the bottom right, the 'Create' button is highlighted with a red box.

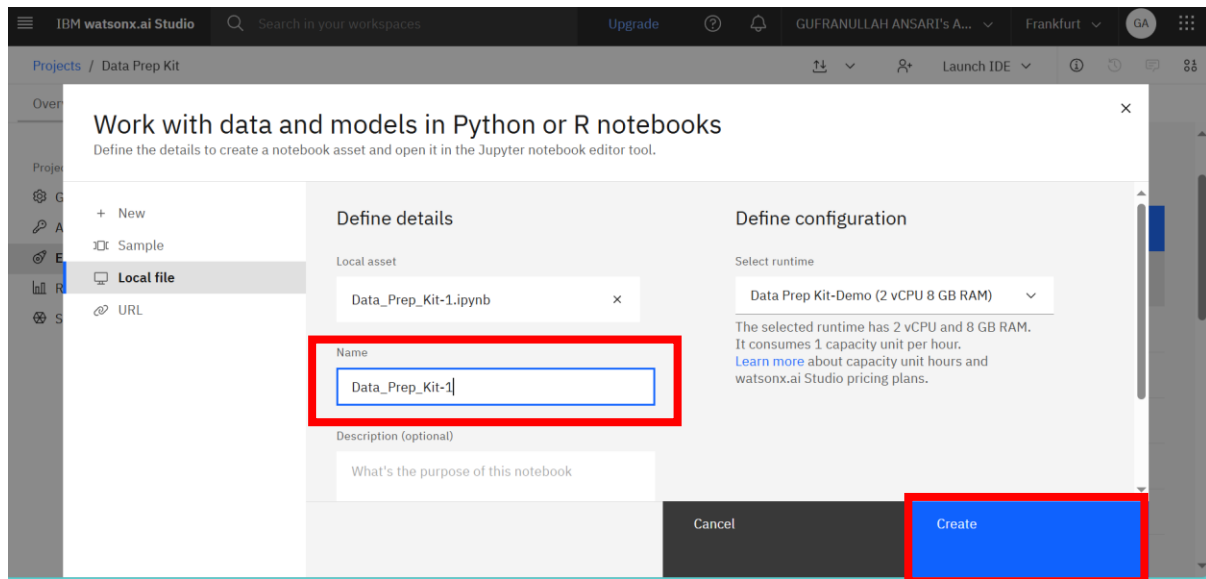
Step 18: Create New notebook



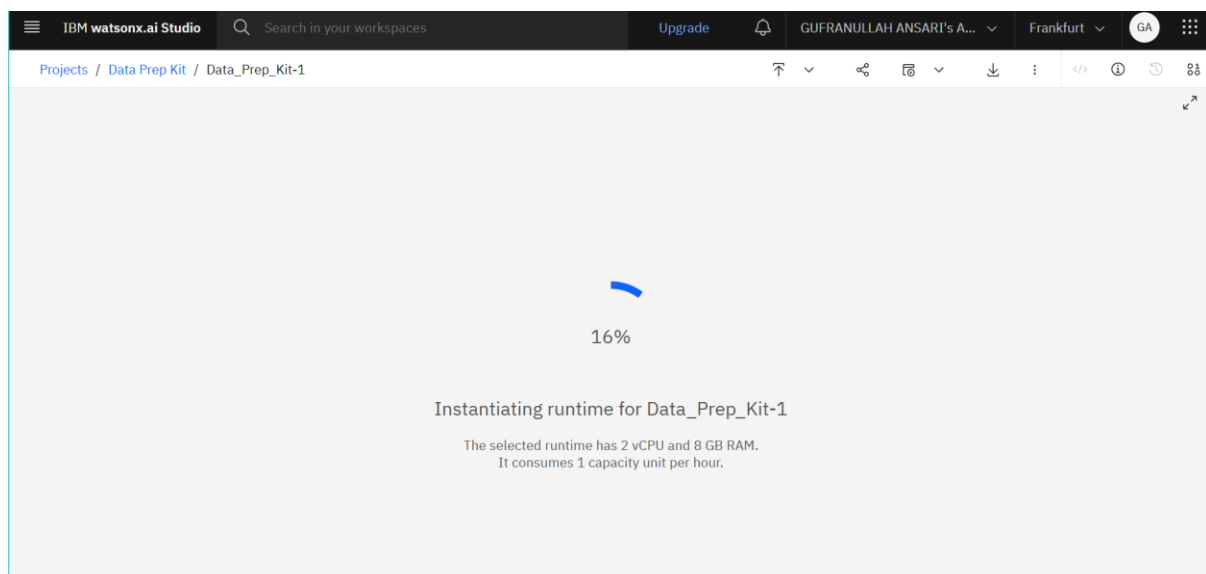
Step 19: Locate file in local drive then browse and upload



Step 20: Define details and create



Step 21: Wait for 100% to complete and load the Notebook.



Step 21: Now Notebook is ready

```
Projects / Data Prep Kit / Data_Prep_Kit-1
File Edit View Run Kernel Help Trusted Memory:2 / 8 GB Python 3.11
[2]: %capture
      !pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2"
      !pip install pandas
      !import urllib.request
      !import shutil

The following notebook example will allow you to test DPK, without cloning the repo. You can run it either on IBM Cloud Jupyter notebook or you can use your local environment (by downloading just the notebook). We use a temporary folder for input and output, but users are encouraged to use their own input folder.
```

Step 22: Run python code

```
[*]: %capture
      !pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2"
      !pip install pandas
      !import urllib.request
      !import shutil

[4]: !shutil.os.makedirs("tmp/input", exist_ok=True)
      !urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/archive1.zip", "tmp/input/archive1.zip")
      !urllib.request.urlretrieve("https://raw.githubusercontent.com/IBM/data-prep-kit/dev/transforms/language/pdf2parquet/test-data/input/redp5110-ch1.pdf", "tmp/input/redp5110-ch1.pdf")

[4]: ('tmp/input/redp5110-ch1.pdf', <http.client.HTTPMessage at 0x7f1f012e150>)

[6]: !pip install --upgrade numpy
      !pip install --upgrade pandas

Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (2.2.4)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.3)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.2.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz in /usr/local/lib/python3.11/dist-packages (from pandas) (2023.3)
Requirement already satisfied: tzdata in /usr/local/lib/python3.11/dist-packages (from pandas) (2023.3)

[7]: !pip install virtualenv
      !virtualenv venv
      !source venv/bin/activate
      !pip install "data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2" pandas
      #Continue with your notebook from here, but ensure you run all installations within the created virtual environment.

Collecting virtualenv
  Downloading virtualenv-20.29.3-py3-none-any.whl.metadata (4.5 kB)
Collecting distlib<1,=>0.3.7 (from virtualenv)
  Downloading distlib-0.3.9-py2.py3-none-any.whl.metadata (5.2 kB)
Requirement already satisfied: filelock<4,>=3.12.2 in /usr/local/lib/python3.11/dist-packages (from virtualenv) (3.18.0)
Requirement already satisfied: platformdirs<5,>=3.9.1 in /usr/local/lib/python3.11/dist-packages (from virtualenv) (4.3.7)
Downloading virtualenv-20.29.3-py3-none-any.whl (4.3 MB)
4.3/4.3 MB 37.3 MB/s eta 0:00:00
Downloading distlib-0.3.9-py2.py3-none-any.whl (468 kB)
469.0/469.0 kB 37.5 MB/s eta 0:00:00
Installing collected packages: distlib, virtualenv
Successfully installed distlib-0.3.9 virtualenv-20.29.3
created virtual environment CPython3.11.11.final.0-64 in 894ms
creator CPython3Posix(dest=/content/venv, clear=False, no_vcs_ignore=False, global=False)
seeder FromAppData(download=False, pip=bundle, setuptools=bundle, wheel=bundle, via=copy, app_data_dir=/root/.local/share/virtualenv)
added seed packages: pip=25.0.1, setuptools=75.8.0, wheel=0.45.1
activators BashActivator,CShellActivator,FishActivator,NushellActivator,PowerShellActivator,PythonActivator
Requirement already satisfied: data-prep-toolkit-transforms==1.0.0a2 in /usr/local/lib/python3.11/dist-packages (from data-prep-toolkit-transforms[pdf2parquet]==1.0.0a2) (1.0.0a2)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.3)

[1]: from dpk_pdf2parquet.transform_python import Pdf2Parquet
      from dpk_pdf2parquet.transform import pdf2parquet_contents_types

[2]: Pdf2Parquet(input_folder="tmp/input",
                output_folder="tmp/output",
                data_files_to_use=['.pdf', '.zip'],
                pdf2parquet_contents_type=pdf2parquet_contents_types.JSON).transform()

09:19:03 INFO - pdf2parquet parameters are : {'batch_size': -1, 'artifacts_path': None, 'contents_type': <pdf2parquet_contents_types.JSON: 'application/json'>, 'do_table_structure': True, 'do_ocr': True, 'ocr_engine': <pdf2parquet_ocr_engine.EASYOCR: 'easyocr'>, 'bitmap_area_threshold': 0.05, 'pdf_backend': <pdf2parquet_pdf_backend.DLPARSE_V2: 'dlparse_v2'>, 'double_precision': 8}
INFO:dpk_pdf2parquet.transform:pdf2parquet parameters are : {'batch_size': -1, 'artifacts_path': None, 'contents_type': <pdf2parquet_contents_types.JSON: 'application/json'>, 'do_table_structure': True, 'do_ocr': True, 'ocr_engine': <pdf2parquet_ocr_engine.EASYOCR: 'easyocr'>, 'bitmap_area_threshold': 0.05, 'pdf_backend': <pdf2parquet_pdf_backend.DLPARSE_V2: 'dlparse_v2'>, 'double_precision': 8}
09:19:03 INFO - pipeline id pipeline_id
INFO:data_processing.runtime.execution_configuration:pipeline id pipeline_id
09:19:03 INFO - code location None
INFO:data_processing.runtime.execution_configuration:code location None
09:19:03 INFO - data factory data_ is using local data access: input_folder - tmp/input output_folder - tmp/output
INFO:data_processing.data_access.data_access_factory_base24f4f209-4dc9-470d-8b10-182b728257bf:data factory data_ is using local data access: input_folder - tmp/input output_folder - tmp/output
09:19:03 INFO - data factory data_max_files -1, n_sample -1
```

```
[3]: import pyarrow.parquet as pq
import pandas as pd
table = pq.read_table('tmp/output/archive1.parquet')
table.to_pandas()
```

	filename	contents	num_pages	num_tables	num_doc_elements	document_id	document_hash	ext
0	2305.03393v1-pg9.pdf	["schema_name":"DoclingDocument","version":"1....	1	1	9	3f8c2d02-fb01-4bc5-8126-ada70eab0296	3463920545297462180	pdf 467dcf637d2efd3f7ad3e
1	2408.09869v1-pg1.pdf	["schema_name":"DoclingDocument","version":"1....	1	0	12	ec7e5fbe-7e73-4cb7-9847-710a67b806d3	582377908831471240	pdf 8ed0cb6d8767bacf9ce8t

```
[4]: table = pq.read_table('tmp/output/redp5110-ch1.parquet')
table.to_pandas()
```

```
[4]: table = pq.read_table('tmp/output/redp5110-ch1.parquet')
table.to_pandas()
```

	filename	contents	num_pages	num_tables	num_doc_elements	document_id	document_hash	ext
0	redp5110-ch1.pdf	["schema_name":"DoclingDocument","version":"1....	5	0	48	04373e22-bd2f-4c78-898c-1e9caf08b18f	74198560999363607	pdf 572c2937fa0e2659f43d74d1576:

Congratulations! You successfully Run Data Prep Kit using IBM Cloud.