



INTERNAL

## SAP Data Intelligence hands-on exercises

This document will guide you step-by-step through the process of training and implementing a text analysis and developing a cluster machine learning model using Python and SAP DI Pipelines.

[www.sap.com/contactsap](http://www.sap.com/contactsap)

© 2018 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. See [www.sap.com/copyright](http://www.sap.com/copyright) for additional trademark information and notices.

## Table of Contents

DISCLAIMER .....	4
OBJECTIVE .....	4
SCENARIO .....	4
ENVIRONMENT ACCESS .....	5
STEP 1 – USE A JUPYTER NOTEBOOK.....	6
STEP 2 – BUILD MODEL PIPELINES .....	9
STEP 3 – USE YOUR CLUSTER MODEL .....	22

## **DISCLAIMER**

The information shared in this document is confidential and proprietary to SAP and may not be disclosed without the permission of SAP. All functionality presented here is subject to change and may be changed by SAP at any time for any reason without notice.

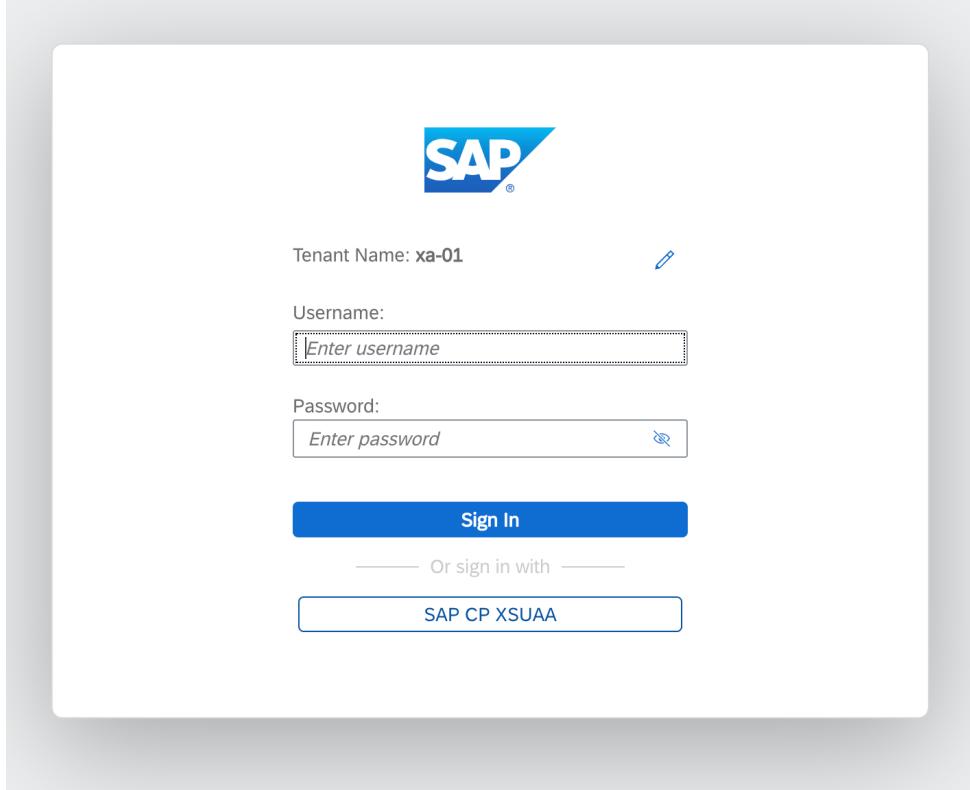
## **OBJECTIVE**

The objective of this exercise is to give you an overview of how you can use the machine learning capabilities in SAP Data Intelligence.

## **SCENARIO**

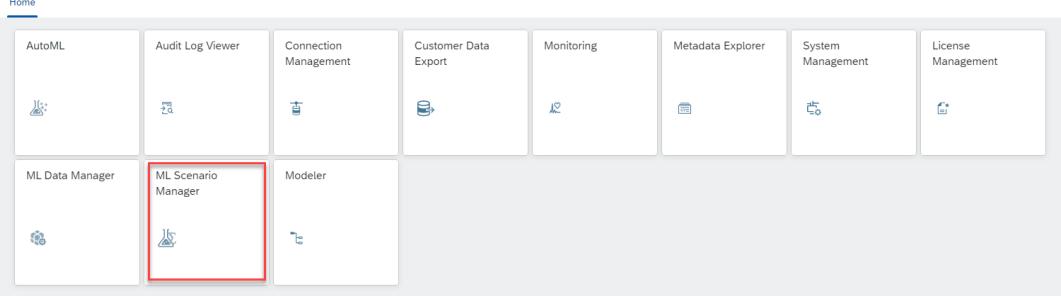
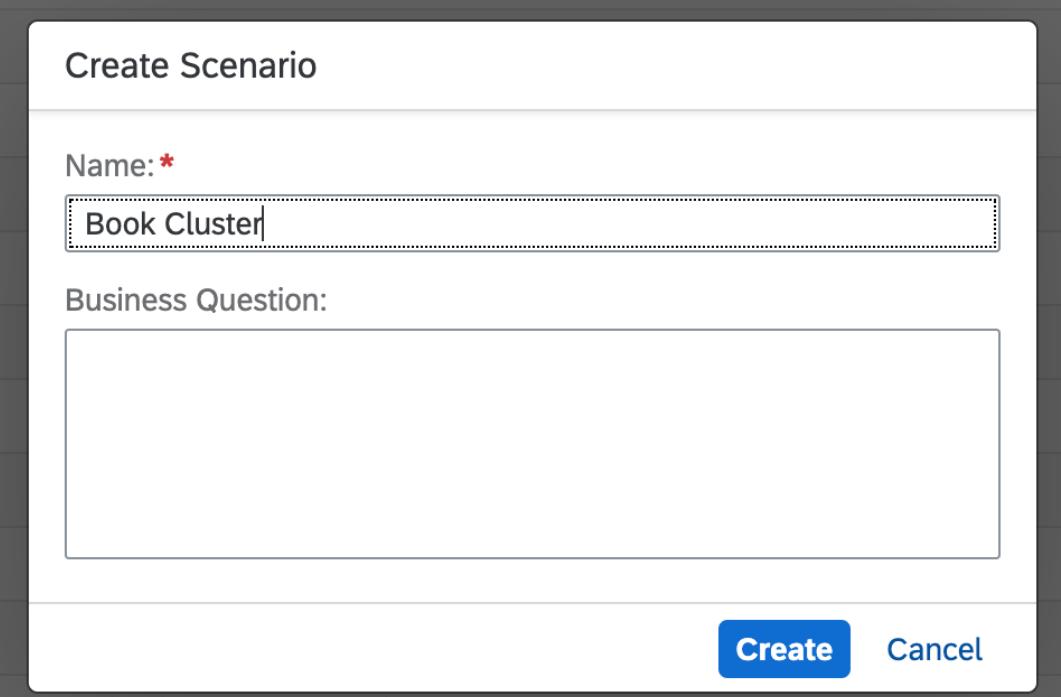
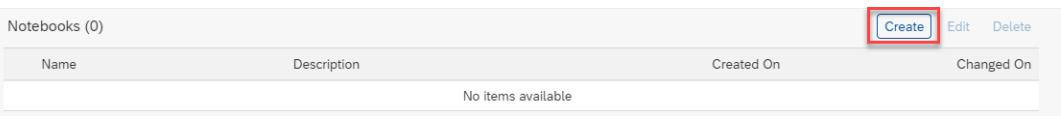
Books are grouped together in a bookshop based on their similarity, so that customers browsing for a book will find lots of similar books on the same shelf. This exercise analyzes the book description data using Python text mining algorithms and then uses this information to assign each book to a cluster. The books within a cluster are as similar as possible (based on the book description), so they are as homogeneous as possible, and there is as wide a difference as possible between clusters, so the different clusters are heterogeneous.

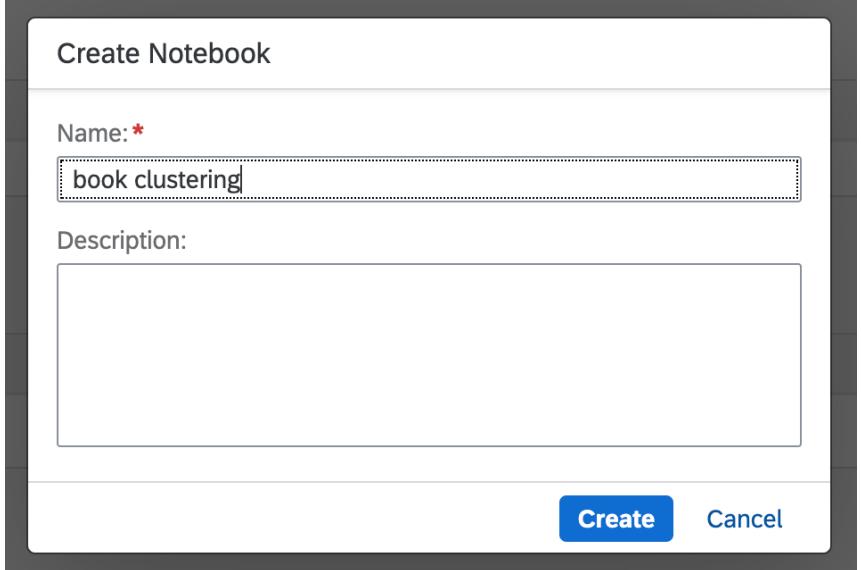
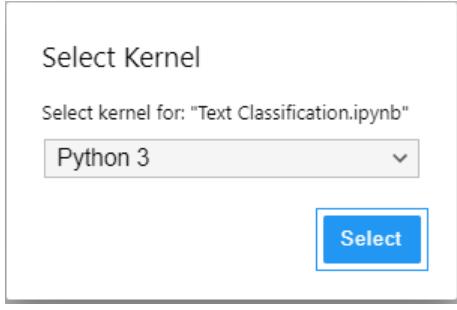
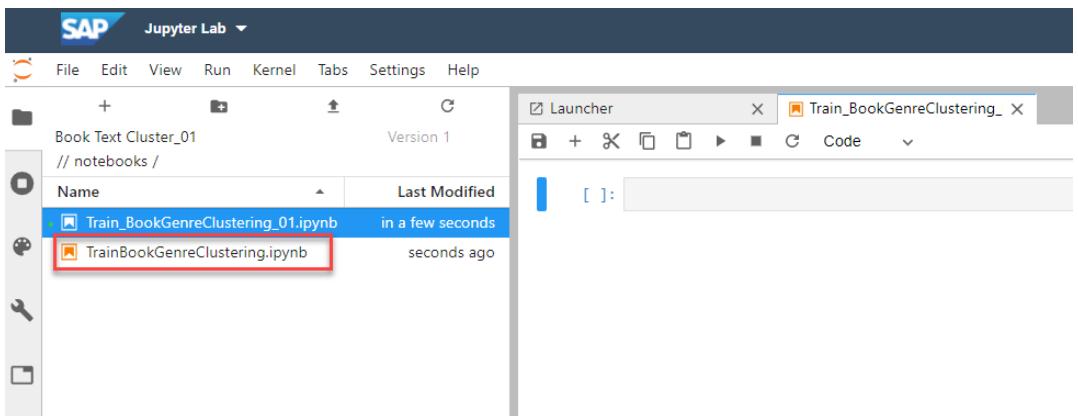
## ENVIRONMENT ACCESS

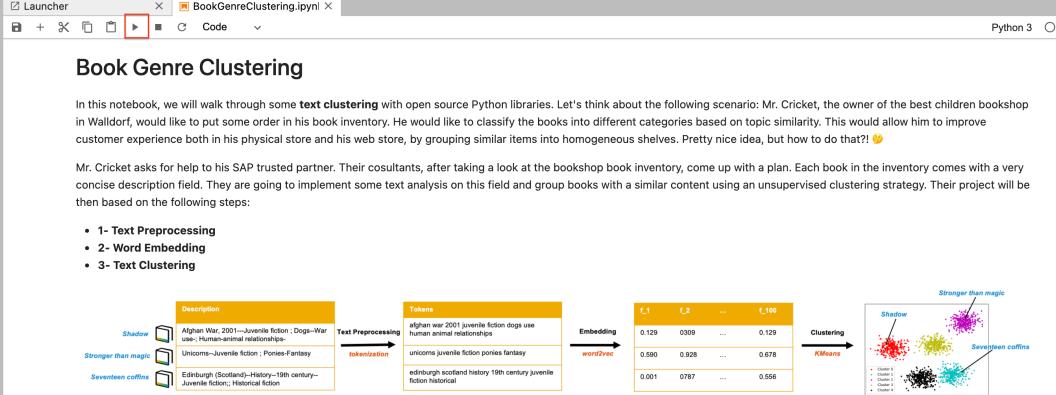
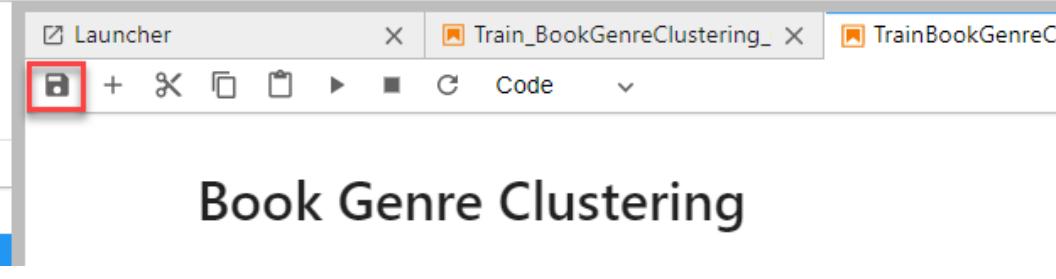
Explanation	Screenshot
Login to your SAP Data Intelligence Cloud tenant with the credentials provided by your instructor	

## STEP 1 – USE A JUPYTER NOTEBOOK

A Jupyter Notebook environment is used to explore the data, and to run predictive model tests to compare the accuracy of different algorithms and the best settings for the hyper-parameters for the algorithms.

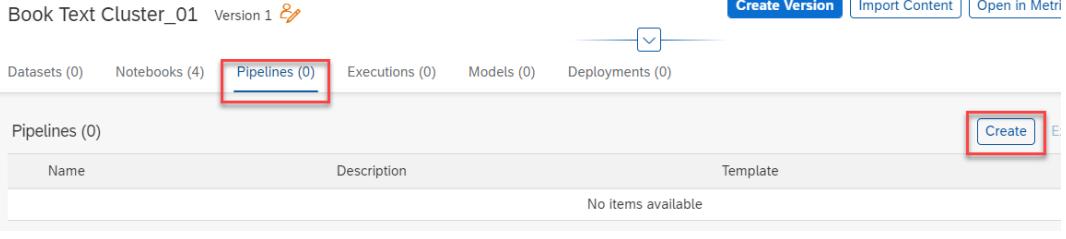
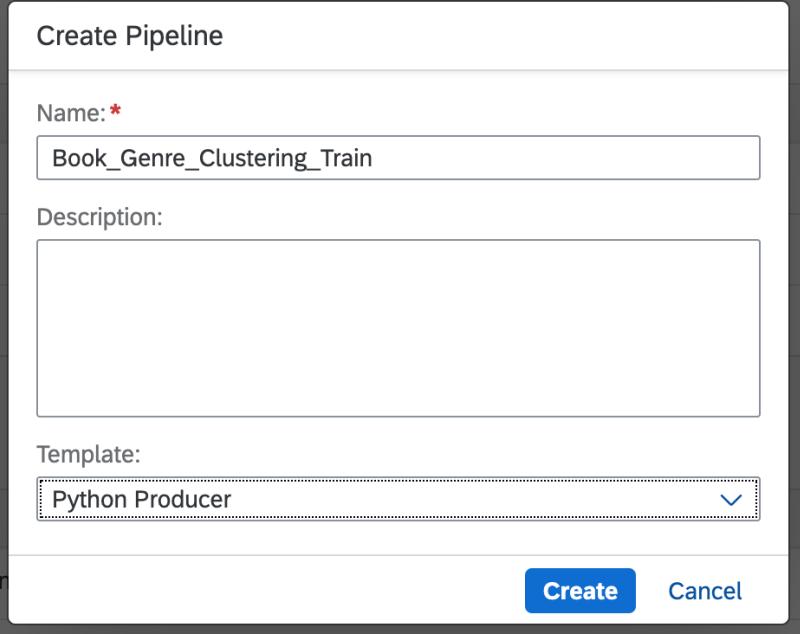
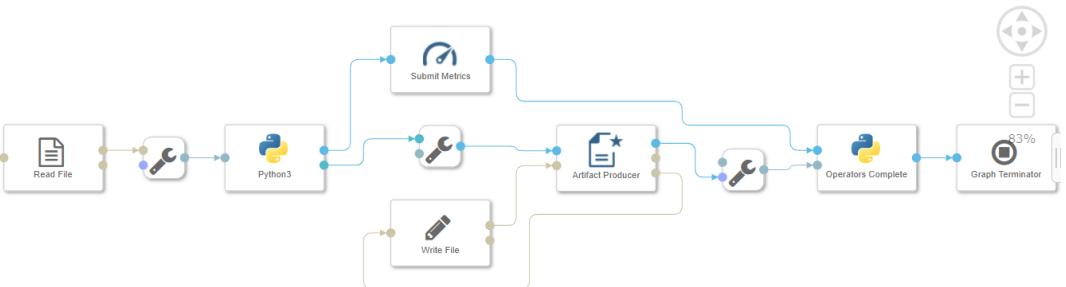
Explanation	Screenshot
Click to open ML Scenario Manager	
<p>Click the Create button. Create a new scenario. Name the scenario “Book Cluster &lt;your user id&gt;”</p> <p>You see the empty scenario. First, you will use the Notebooks to explore the data and to script the text analysis and cluster model in Python. Next, pipelines bring the code into production. Executions of these pipelines will create Machine Learning models, which are then deployed as REST-API for inference.</p>	
In the Notebooks section, click Create to create a new notebook.	

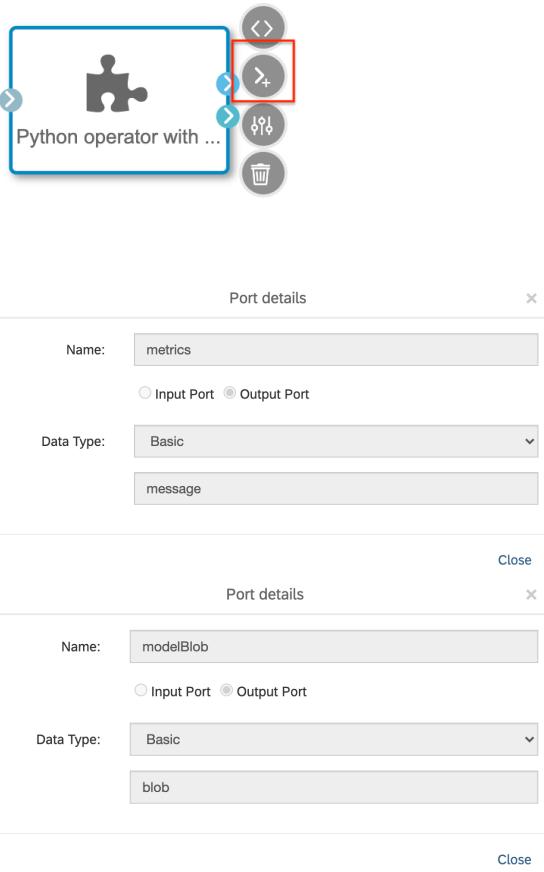
Explanation	Screenshot
Name the notebook "book clustering"	 <p>Create Notebook</p> <p>Name: *</p> <input type="text" value="book clustering"/> <p>Description:</p> <p><b>Create</b>   <b>Cancel</b></p>
Select the Python 3 Kernel option.	 <p>Select Kernel</p> <p>Select kernel for: "Text Classification.ipynb"</p> <p>Python 3</p> <p><b>Select</b></p>
<p>Use the Upload Files function to upload the Python code into the console.</p> <p>Upload file <b>BookGenreClustering.ipynb</b> and <b>text_clustering.png</b> (you can find them in the bookcamp github repository)</p> <p>Double click on the notebook that is uploaded.</p>	 <p>SAP Jupyter Lab</p> <p>File Edit View Run Kernel Tabs Settings Help</p> <p>Book Text Cluster_01 Version 1</p> <p>Name Last Modified</p> <p>TrainBookGenreClustering.ipynb in a few seconds</p> <p>TrainBookGenreClustering.ipynb seconds ago</p> <p>Launcher Train_BookGenreClustering_ X</p> <p>Select file BookGenreClustering.ipynb</p>

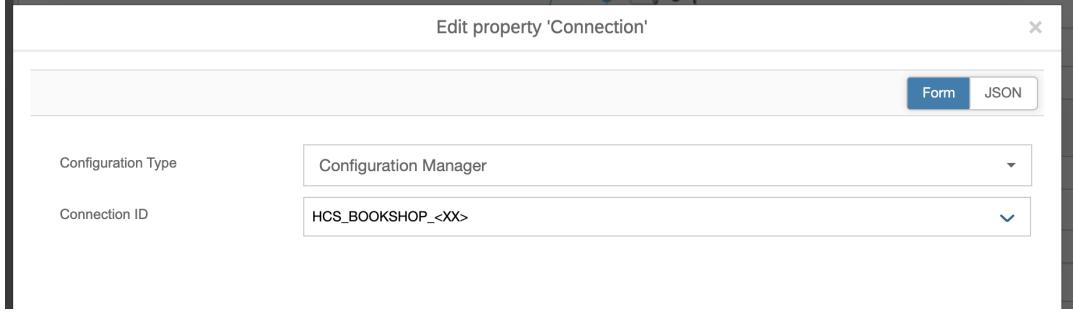
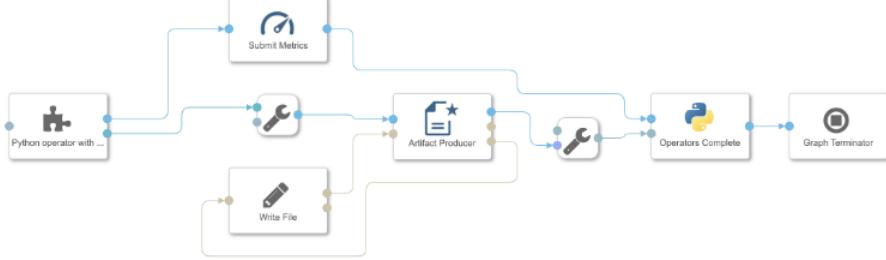
Explanation	Screenshot
<p>Double click on the notebook that is uploaded.</p> <p>Step through the code to check it works correctly.</p> <p>Highlight each cell in sequence and click the arrow button to run the selected cell and advance. Note that line 9 loads the word embedding data and line 19 runs a tsne model. Both of these steps will take a few minutes to complete. Analyze the results of your exploratory data analysis and understand the text analysis and cluster model results.</p>	 <p>The screenshot shows a Jupyter Notebook interface with the title "Book Genre Clustering". The notebook content includes a flow diagram illustrating the text processing pipeline: "Description" leads to "Text Preprocessing" (Tokenization), which leads to "Embedding" (word2vec). This is followed by a t-SNE clustering visualization showing four distinct clusters: "Shadow", "Stronger than magic", "Unicorn", and "Seventeen coffins". Below the diagram, a note says "Let's put this into practice!" and lists required Python libraries: pandas, numpy, matplotlib, seaborn, regex, and nltk. A code cell [1] contains the following Python code:</p> <pre>[1]: # basic Python !pip install pandas !pip install numpy !pip install matplotlib !pip install seaborn  # text preprocessing !pip install regex !pip install nltk</pre>
<p>Once you have stepped through to the end of the notebook, and there are no errors, save the notebook.</p>	

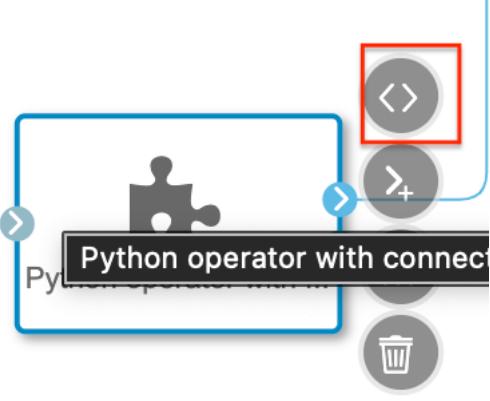
## STEP 2 – BUILD MODEL PIPELINES

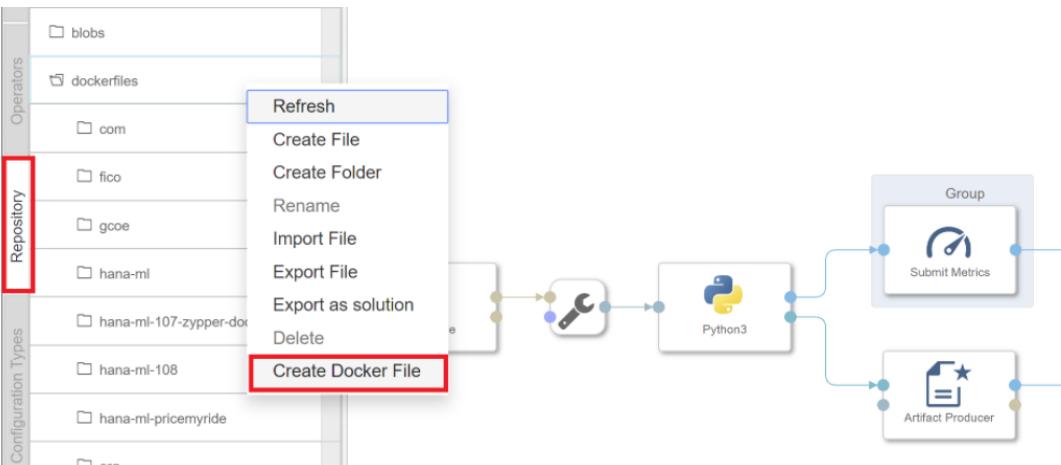
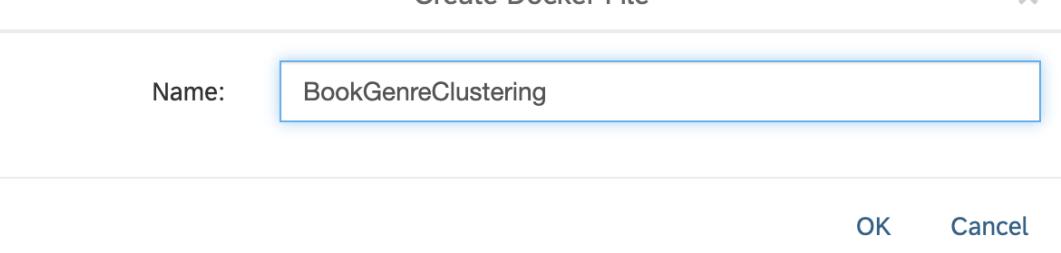
Model pipelines are used to operationalize the cluster model. Now that you have prepared the data, identify the best algorithm to use and which hyper-parameters work best, you want to take the model to production. You will build two pipeline. The first one is used to automate the model training, while the second one is used to inference the model on new data.

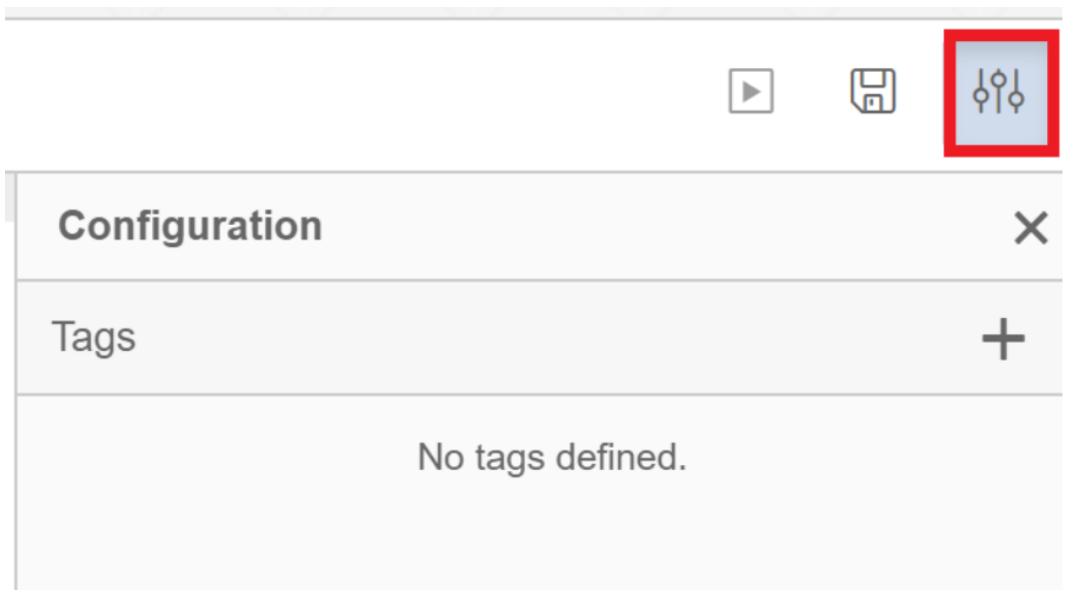
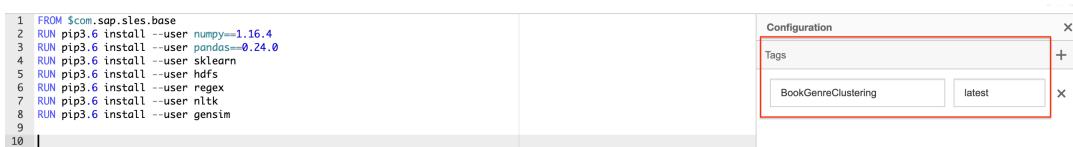
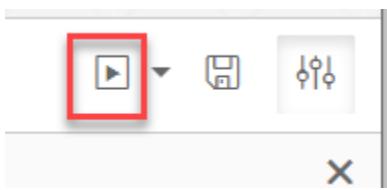
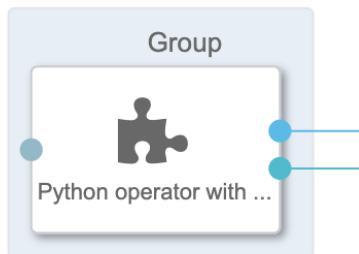
Explanation	Screenshot
<p>To create the graphical pipeline to retrain the model, go to your ML Scenario's main page, select the “Pipelines” tab and click Create.</p>	 <p>The screenshot shows the 'Book Text Cluster_01 Version 1' interface. At the top, there are tabs for Datasets (0), Notebooks (4), Pipelines (0) (which is highlighted with a red box), Executions (0), Models (0), and Deployments (0). Below the tabs, there is a table titled 'Pipelines (0)' with columns for Name, Description, and Template. A large red box highlights the 'Create' button in the top right corner of the pipeline list area.</p>
<p>Name the pipeline “Text Clustering Train” and select the “Python Producer” template. Click Create.</p>	 <p>The screenshot shows the 'Create Pipeline' dialog box. It has fields for 'Name:' (set to 'Book_Genre_Clustering_Train') and 'Description:' (empty). Under 'Template:', it shows 'Python Producer' selected. At the bottom are 'Create' and 'Cancel' buttons, with a red box highlighting the 'Create' button.</p>
<p>You need to adjust the pipeline template. The pipeline loads data with the “Read File” operator. The data is passed to a Python operator, where the ML model is trained. The same Python-operator stores the model in the ML</p>	 <p>The screenshot shows the completed ML Pipeline graph. The flow starts with a 'Read File' operator, followed by a 'Python3' operator (where the ML model is trained). This is followed by a 'Submit Metrics' operator, another 'Python3' operator, an 'Artifact Producer' operator, a 'Write File' operator, an 'Operators Complete' operator, and finally a 'Graph Terminator' operator. There are also intermediate operators like a lock operator and a '83%' progress bar. A red box highlights the 'Graph Terminator' operator.</p>

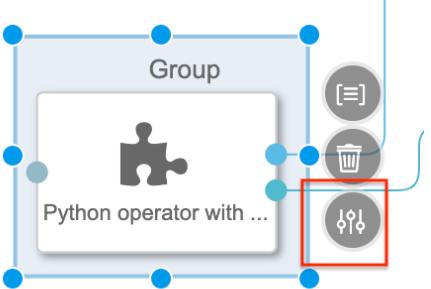
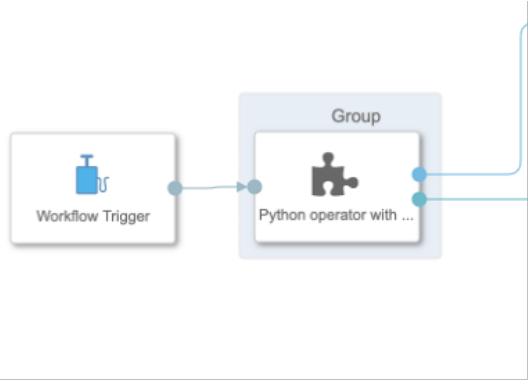
Explanation	Screenshot
<p>Scenario through the “Artifact Producer”. The Python-operator’s second output passes a model quality metric to the ML Scenario. Once both model and metric are saved, the pipeline’s execution is ended with the “Graph Terminator”.</p>	
<p>Since for us the input data are not stored in a file, but in HANA Cloud, we don’t need the Read File operator. We will adapt the custom python operator we built in the previous exercise.</p> <p>Insert a Python with HANA ML connection operator in the canvas. Click on the Add port symbol to add two output ports.</p> <p>Configure the ports with the parameter shown in the picture</p>	 <p>The screenshot shows the Data Integration interface. A Python operator node is selected, indicated by a blue border. To its right, there are several icons: a double-headed arrow, a plus sign inside a circle (highlighted with a red box), a gear, and a trash can. Below the node, two 'Port details' dialogs are displayed. The top dialog is for 'metrics', set as an 'Output Port' of type 'Basic' message. The bottom dialog is for 'modelBlob', also set as an 'Output Port' of type 'Basic' blob.</p>

Explanation	Screenshot
<p>In the Configuration panel, open the “Connection” configuration, set “Configuration type” to “Connection Management” and select from the drop-down menu the <b>connection ID</b> related to your Hana Cloud instance. You should use the connection created in DV150_ex2 if the exercise were required otherwise the predefined connection. Check the exercises notes in the Teams channel for the connection to use.</p>	
<p>Delete the Read File operator, and the Conversion operator. Replace the Python Operator with the custom operator that you have just configured your pipeline should look like in the picture.</p>	

Explanation	Screenshot
<p>Next, adjust the Python code that analyzes the text data and trains the cluster model.</p> <p>Select the custom operator and click the Script option.</p>	
<p>The template code opens up. It shows how to pass the text analysis, model and metrics into the ML Scenario.</p> <p><b>Carefully</b> copy and paste to replace the existing code with the code given here, so that we can operationalize the clustering model we developed in the notebook.</p>	<p><a href="https://github.com/SAP-samples/btp-data-to-value-workshop/blob/main/02-data-modeling%26processing/exercises/code_snippets/dv220_train.py">https://github.com/SAP-samples/btp-data-to-value-workshop/blob/main/02-data-modeling%26processing/exercises/code_snippets/dv220_train.py</a></p>
<p>Close the Script-window, then click "Save" in the menu bar.</p>	

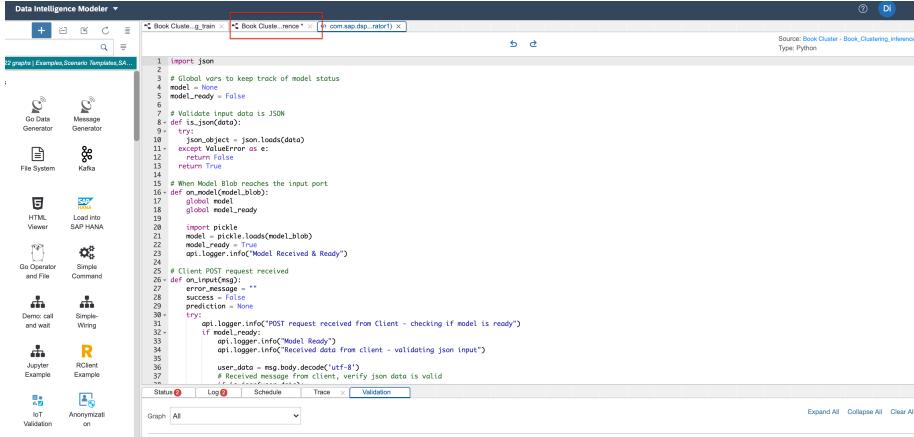
Explanation	Screenshot
<p>You need to create a Docker image for the custom python operator. This gives the flexibility to leverage virtually any Python library. The Docker file installs the necessary libraries. You find the docker files by clicking into the “Repository” tab on the left, then right-click the “<b>dockerfiles</b>” folder and select “Create Docker File”.</p>	
<p>Name the file BookGenreClustering</p>	
<p>Enter this code into the Docker File window. This code leverages a base image that comes with SAP Data Intelligence and installs the necessary libraries on it.</p>	<pre data-bbox="437 1262 833 1537">FROM \$com.sap.sles.base RUN pip install --user numpy RUN pip install --user pandas RUN pip install --user sklearn RUN pip install --user hana_ml RUN pip install --user regex RUN pip install --user nltk RUN pip install --user gensim</pre>

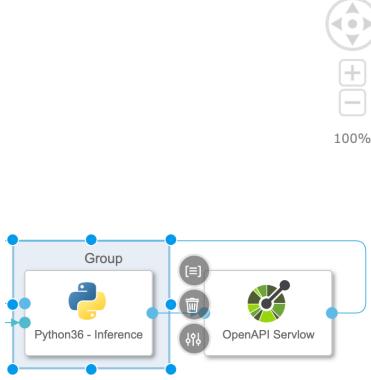
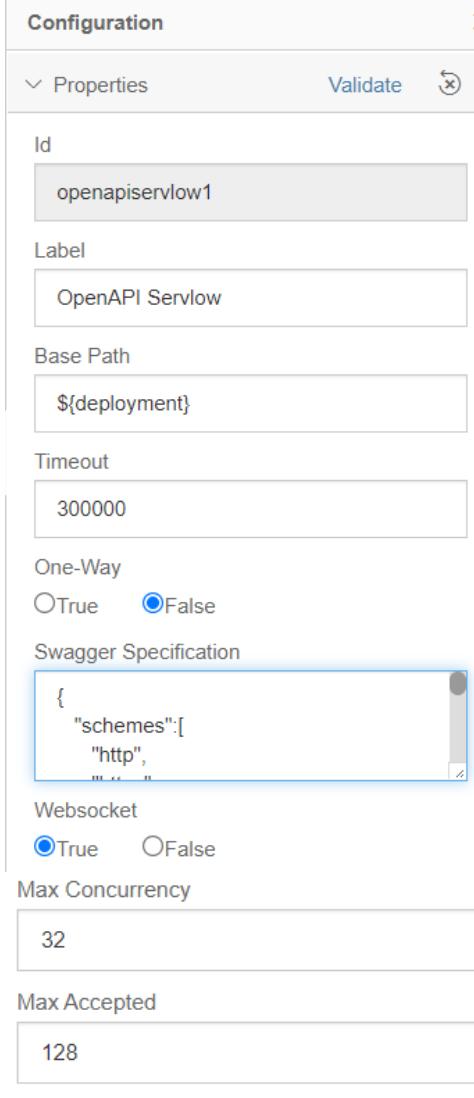
Explanation	Screenshot
<p>Open the Configuration panel for the Docker File with the icon on the top-right hand corner.</p>	
<p>Assign a tag to the docker image</p>	
<p>Now save the Docker file and click the “Build” icon to start building the Docker image. Wait a few minutes and you should receive a confirmation that the build completed successfully.</p>	
<p>Now you need to configure the custom operator, which trains the model, to use this Docker image. Go back to the graphical pipeline and right-click the custom operator and select “Group”.</p>	

Explanation	Screenshot
<p>You specify which Docker image should be used. Select the group which surrounds the “Python 3” Operator. In the group’s Configuration select the docker image you have built.</p>	 <div data-bbox="430 599 1491 1417"> <p><b>Configuration</b> <span style="float: right;">X</span></p> <p><b>Properties</b> <span style="float: right;">Validate</span></p> <p><b>Id</b> group1</p> <p><b>Description</b> Group</p> <p><b>Restart Policy</b></p> <div style="border: 1px solid black; padding: 2px;"> <b>Name:</b> metricsResponse  <b>Data Type:</b> message       </div> <p><b>Tags</b> <span style="float: right;">+</span>  <input type="text" value="BookGenreClustering"/> <span style="float: right;">▼</span> <input type="text" value="latest"/> <span style="float: right;">▼</span> <span style="float: right;">X</span></p> <p><b>Multiplicity</b></p> </div>
<p>Select the Workflow Trigger operator in the operators panel. Add it to the pipeline and connect it to the trigger port of the custom operator. Save your pipeline</p>	

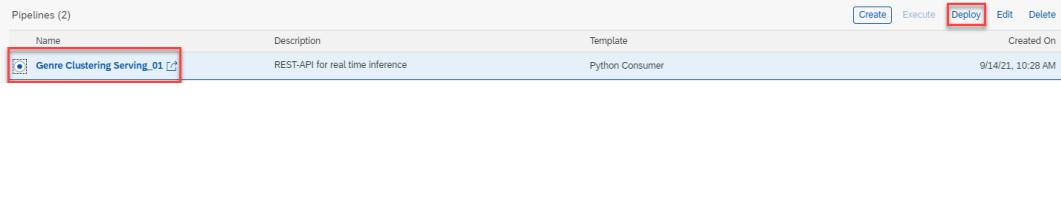
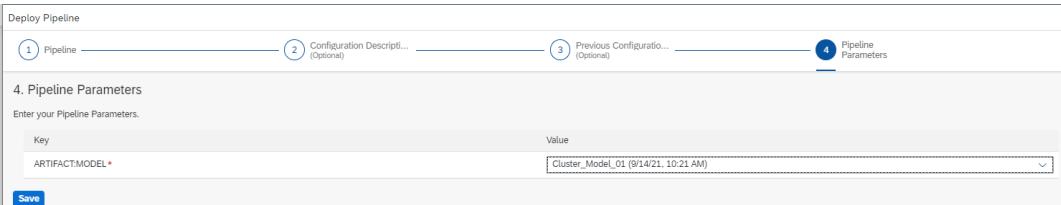
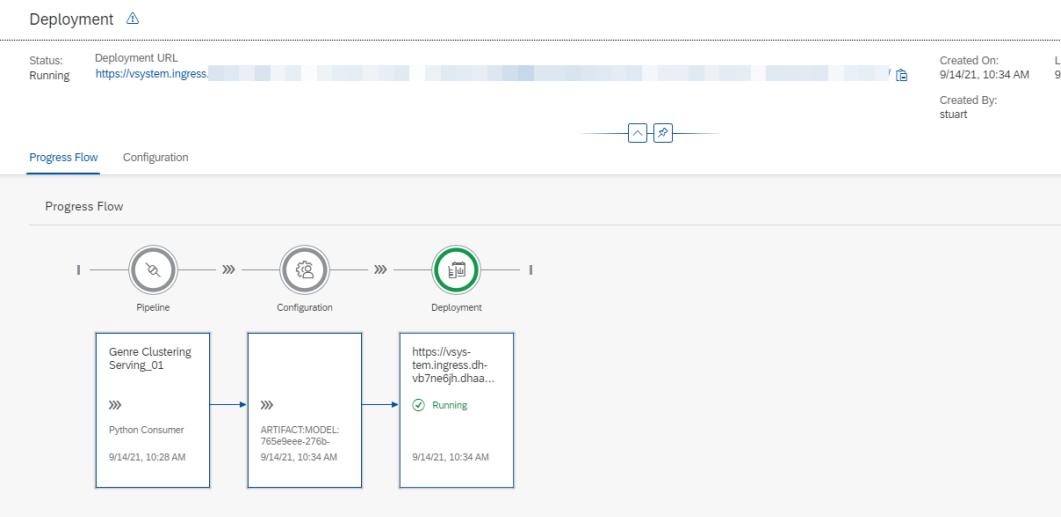
Explanation	Screenshot				
<p>The pipeline is now complete and you can run it. Go back to the ML Scenario. Select the pipeline in the ML Scenario and click the “Execute” button on the right.</p>					
<p>Click “Step 3” and then “Step 4” to skip the optional steps until you get to the “Enter your Pipeline Parameters”. Set “newArtifactName” Value to “kmeans”. Click Save. The trained model will be saved under this name.</p>	<p>4. Pipeline Parameters</p> <p>Enter your Pipeline Parameters.</p> <table border="1" data-bbox="437 656 1491 762"> <thead> <tr> <th data-bbox="437 656 1328 692">Key</th> <th data-bbox="1328 656 1491 692">Value</th> </tr> </thead> <tbody> <tr> <td data-bbox="437 692 1328 728">newArtifactName*</td> <td data-bbox="1328 692 1491 728">kmeans</td> </tr> </tbody> </table> <p><b>Save</b></p>	Key	Value	newArtifactName*	kmeans
Key	Value				
newArtifactName*	kmeans				
<p>Wait a few minutes until the pipeline executes and completes. The metrics section shows the trained model’s silhouette metric. The model itself was saved successfully under the name “kmeans”. The model has a Technical Identifier.</p>	<p>Status: <b>Completed</b>    Created On: 9/14/21, 10:19 AM    Last Synchronized On: 9/14/21, 10:22 AM Created By: stuart</p> <p>Progress Flow   Configuration   Metrics (1)   Models and Datasets (1)</p> <p><b>Progress Flow</b></p> <pre> graph LR     Pipeline[Pipeline] --&gt; Configuration[Configuration]     Configuration --&gt; Execution[Execution]     Execution --&gt; Metrics[Metrics]     Metrics --&gt; Model[Models and Datasets]     </pre> <p>Pipeline: Text Clustering Train_00 Python Producer: 9/14/21, 10:06 AM</p> <p>Configuration: newArtifactName: Cluster_Model_01 9/14/21, 10:19 AM</p> <p>Execution: Execution 9/14/21, 10:19 AM Completed</p> <p>Metrics: SILHOUETTE: 0.04 start: 2021-09-14T10:19:00Z 9/14/21, 10:19 AM</p> <p>Models and Datasets: Cluster_Model_01 9/14/21, 10:21 AM</p>				

Explanation	Screenshot
<p>You will now use the model for real-time inference with REST-API. Go back to the main page of your ML Scenario and create a second pipeline. This pipeline will provide the REST-API to obtain predictions in real-time. Name the pipeline “Book Clustering Consumer”. Select the template “Python Consumer”. This template contains a pipeline that provides a REST-API.</p>	<p>Create Pipeline</p> <p>Name: * <input type="text" value="Book_Clustering_inference"/></p> <p>Description: <input type="text"/></p> <p>Template: <input type="text" value="Python Consumer"/></p> <p><b>Create</b> <b>Cancel</b></p>
<p>The “OpenAPI Servlow” operator provides the REST-API. The “Artifact Consumer” loads the trained model from your ML scenario. The “Python36 – Inference” operator ties the two operators together. It receives the input from the REST-API call (here the user’s text input book description) and uses the loaded model to assign the cluster, which is then returned by the “OpenAPI Servlow” to the client, which</p>	<pre> graph LR     A[Submit Artifact Name] --&gt; B[Artifact Consumer]     B --&gt; C[Read File]     C --&gt; D["Python36 - Inference"]     D --&gt; E[OpenAPI Servlow]   </pre>

Explanation	Screenshot
had called the REST-API.	
You only need to update the script of the Python3.6 Inference operator. Select the Python3.6 operator in the pipeline and click on the script icon	
<b>Carefully copy and paste to replace the whole code with the code given here:</b>	<p><a href="https://github.com/SAP-samples/btp-data-to-value-workshop/blob/main/02-data-modeling%26processing/exercises/code_snippets/dv220_consumer.py">https://github.com/SAP-samples/btp-data-to-value-workshop/blob/main/02-data-modeling%26processing/exercises/code_snippets/dv220_consumer.py</a></p>
Close the editor window.	
Go back to the pipeline tab and save the pipeline	 <pre> 1 import json 2 # Global vars to keep track of model status 3 model = None 4 model_ready = False 5 6 # Validate input data is JSON 7 def is_json(data): 8     try: 9         json_object = json.loads(data) 10    except ValueError as e: 11        return False 12    return True 13 14 # When Model Blob receives the input port 15 def on_model_blob(port): 16     global model 17     global model_ready 18 19     import pickle 20     model = pickle.loads(model_blob) 21     model_ready = True 22     api_logger.info("Model Received &amp; Ready") 23 24 # Client POST request received 25 def on_post_request(port): 26     error_message = '' 27     success = False 28 29     try: 30         # Check if Model is ready 31         if model_ready: 32             api_logger.info("POST request received from Client - checking if model is ready") 33             api_logger.info("Model Ready") 34             api_logger.info("Received data from Client - validating json input") 35 36             user_data = msg_body.decode('utf-8') 37             user_data = json.loads(user_data) 38             # Validate received message from client, verify json data is valid 39 40     except Exception as e: 41         error_message = str(e) 42 43     return success, error_message </pre>

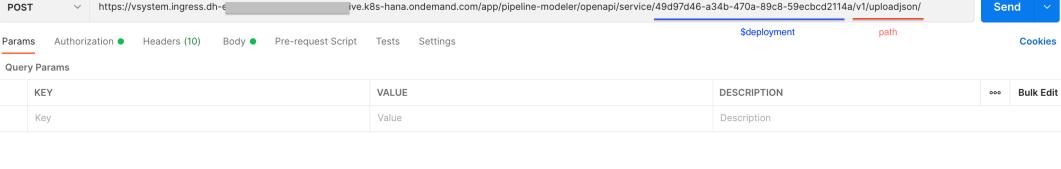
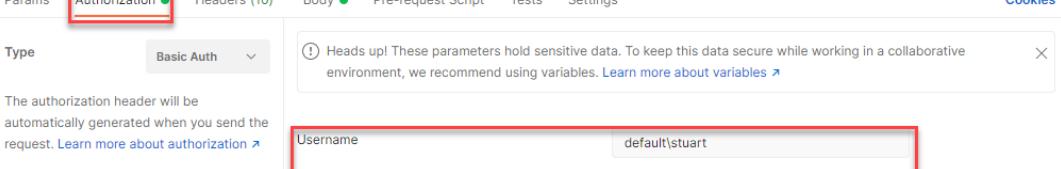
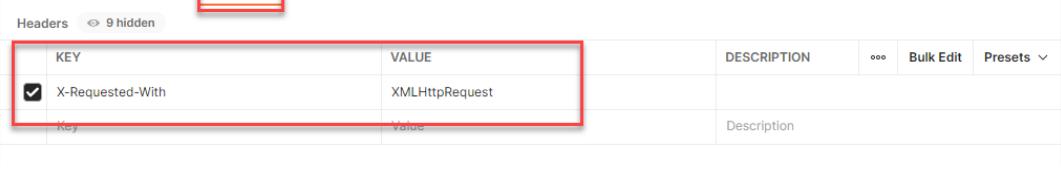
Explanation	Screenshot
<p>Finally, you need to assign the Docker image to the “Python36 – Inference” operator. As before, right-click the operator and select “Group”. Add the tag of your docker image Save the changes.</p>	 <div data-bbox="801 255 1269 762"> <p><b>Configuration</b></p> <p>Properties</p> <p><b>Id</b>: group1</p> <p><b>Description</b>: Group</p> <p><b>Restart Policy</b>: (dropdown menu)</p> <p><b>Tags</b>: BookGenreClustering latest</p> <p><b>Multiplicity</b>: (dropdown menu)</p> <p><b>Resources</b>: (link)</p> </div>
<p>Click on OpenAPIServlow and have a look at the configuration</p>	 <div data-bbox="425 819 899 1917"> <p><b>Configuration</b></p> <p>Properties</p> <p><b>Id</b>: openapiservlow1</p> <p><b>Label</b>: OpenAPI Servlow</p> <p><b>Base Path</b>: \${deployment}</p> <p><b>Timeout</b>: 300000</p> <p><b>One-Way</b>: <input type="radio"/> True <input checked="" type="radio"/> False</p> <p><b>Swagger Specification</b>: (text area with JSON code)</p> <pre>{   "schemes": [     "http",     ...   ] }</pre> <p><b>Websocket</b>: <input checked="" type="radio"/> True <input type="radio"/> False</p> <p><b>Max Concurrency</b>: 32</p> <p><b>Max Accepted</b>: 128</p> </div>

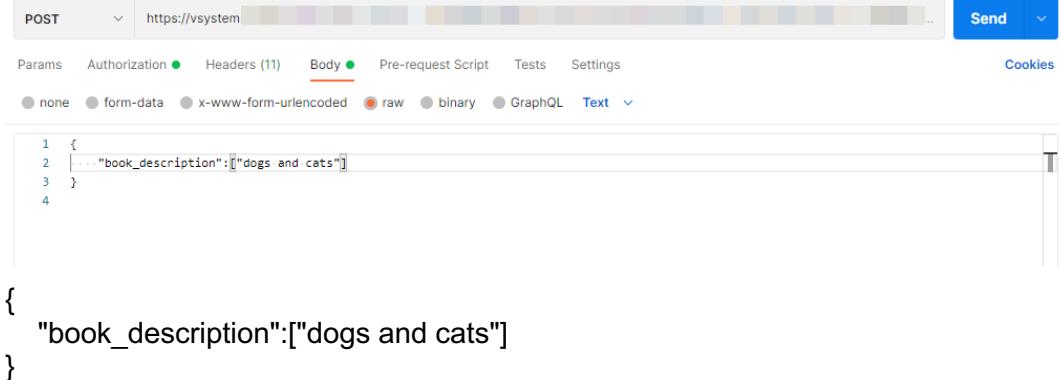
Explanation	Screenshot
<p>Notice in particular the content of the Swagger Specification, <b>but don't change anything here!</b></p>	<pre>{   "schemes":[     "http",     "https"   ],   "swagger":"2.0",   "info":{      "description":"This is an example of using the OpenAPI Servlow to carry out inference with an existing model.",     "title":"OpenAPI demo",     "termsOfService":"http://www.sap.com/vora/terms/",     "contact":{      },     "license":{        "name":"Apache 2.0",       "url":"http://www.apache.org/licenses/LICENSE-2.0.html"     },     "version":"1.0.0"   },   "basePath":"/\$deployment",   "paths":{      "/v1/uploadjson":{        "post":{          "description":"Upload data in json format",         "consumes":[           "application/json"         ],         "produces":[           "application/json"         ],         "summary":"Upload JSON data to be used in the Python operator's script",         "operationId":"upload",         "parameters":[           {              "type":"object",             "description":"json data",             "name":"body",             "in":"body",             "required":true           }         ],         "responses":{            "200":{              "description":"Data uploaded"           },           "500":{              "description":"Error during upload of json"           }         }       }     }   },   "definitions":{    } }</pre>

Explanation	Screenshot
	<pre data-bbox="430 255 703 466"> }, "securityDefinitions":{   "UserSecurity":{     "type":"basic"   } } }</pre>
<p>Go back to the ML Scenario. Now deploy the new pipeline. Select the pipeline and click “Deploy”.</p>	
<p>Click through the screens until you can select the trained model from the drop-down. Click “Save”.</p>	
<p>After a few minutes the pipeline will start running.</p>	

## STEP 3 – USE YOUR CLUSTER MODEL

Now that you have deployed your model, you can use it for real-time cluster assignment. For this, you are going to use the Postman application.

Explanation	Screenshot
<p>Open Postman. Copy the deployment URL from SAP DI. Enter the Deployment URL as request URL. Extend the URL with <b>v1/uploadjson/</b>, the path specified in the OpenAPI servlow operator. Change the request type from “GET” to “POST”.</p>	<p>Copy Deployment URL in SAP DI:</p>  <p>To Postman:</p> 
<p>Go to the “Authorization” tab. Select “Basic Auth” and enter your username and password for SAP Data Intelligence. The username starts with your tenant’s name, followed by a backslash and your actual username.</p>	
<p>Go to the “Headers” tab and enter the key “X-Requested-With” with value “XMLHttpRequest”.</p>	

Explanation	Screenshot
<p>Finally, pass the input data to the REST-API. Select the “Body” tab, choose “raw” and enter the syntax given here.</p> <pre data-bbox="425 255 1486 635">{   "book_description":["dogs and cats"] }</pre>	
<p>Press “Send” and after a few minutes you will see the genre prediction that comes from SAP Data Intelligence. Try the REST-API with different text to see how the cluster allocations change.</p>	 <pre data-bbox="474 846 768 1015">1  { 2    "Cluster": [ 3      9 4    ] 5  }</pre>
<p>You have now completed the exercise.</p>	

