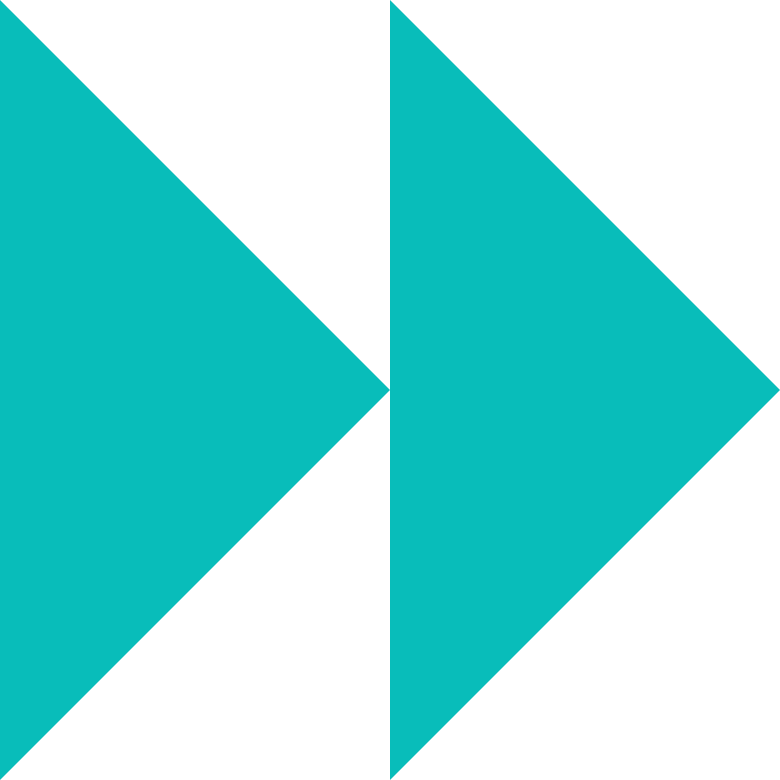
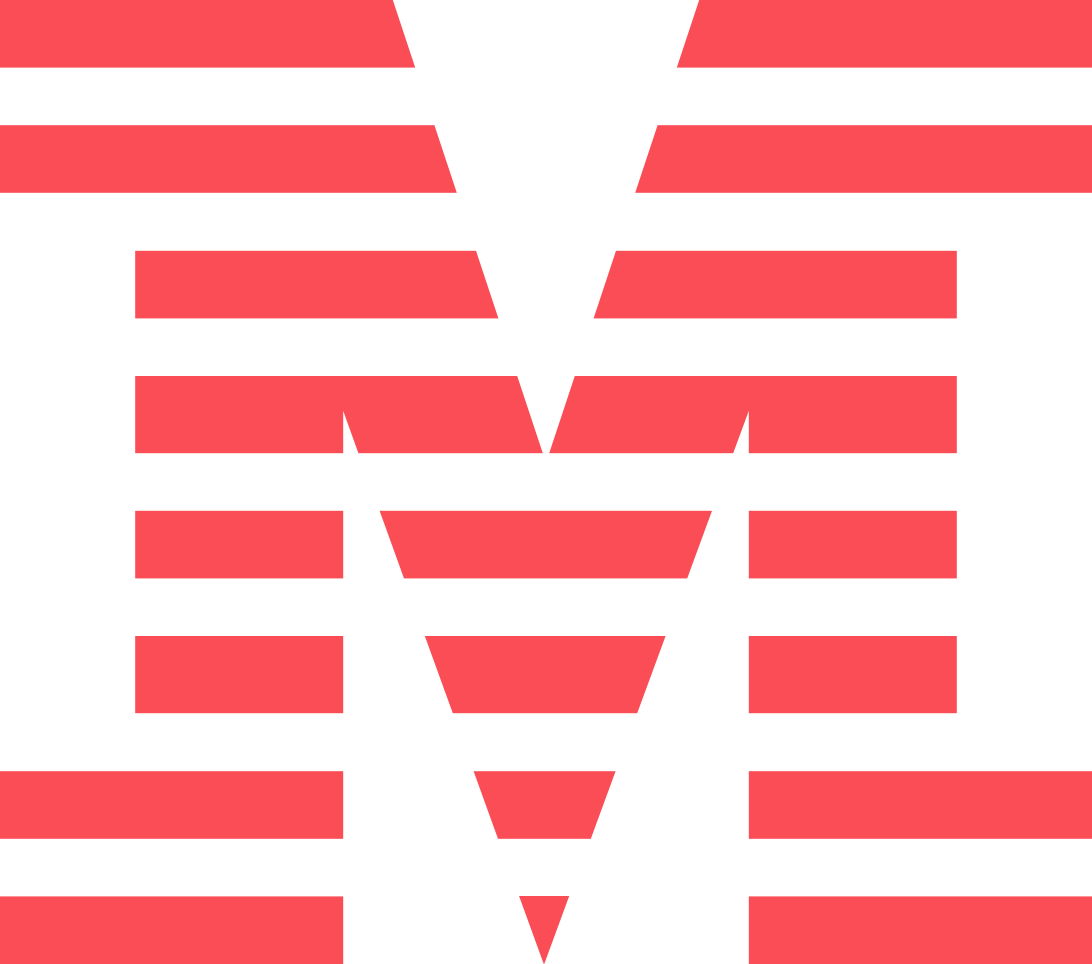
Citing your sources: Explaining

generative AI output

Hands-on lab guide

Eric Martens  
[emartens@us.ibm.com](mailto:emartens@us.ibm.com)  
Learning Content Development, Data and AI



Loucas Loumakos  
[loucas.loumakos@ibm.com](mailto:loucas.loumakos@ibm.com)  
Learning Content Development, Data and AI

Table of Contents

[1 Introduction 5](#_Toc175866258)

[1.1 About this hands-on lab 5](#_Toc175866259)

[2 Setting up the environment 5](#_Toc175866260)

[2.1 Create a watsonx project 5](#_Toc175866261)

[2.2 Associate a machine learning service 8](#_Toc175866262)

[3 Executing the Jupyter notebook 9](#_Toc175866263)

[3.1 Sub task – Use to break up major tasks into smaller sets of steps **Error! Bookmark not defined.**](#_Toc175866264)

[3.2 Sub task – Use to break up major tasks into smaller sets of steps **Error! Bookmark not defined.**](#_Toc175866265)

[4 Integrating with an application 10](#_Toc175866266)

[4.1 Creating the virtual assistant **Error! Bookmark not defined.**](#_Toc175866267)

[4.2 Adding search skills to the assistant **Error! Bookmark not defined.**](#_Toc175866268)

[4.3 Testing the assistant **Error! Bookmark not defined.**](#_Toc175866269)

[5 Getting help and troubleshooting 10](#_Toc175866270)

[5.1 Common troubleshooting tips 10](#_Toc175866271)

[5.2 Getting help 11](#_Toc175866272)

Notices and disclaimers

© 2024 International Business Machines Corporation. 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.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

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 per the terms and conditions of the agreements under which they are provided. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Notices and disclaimers (Continued)

It is the customer’s responsibility to ensure 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 follows any law.

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 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, and ibm.com 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

[Learn more →](http://www.ibm.com/legal/copytrade.shtml)

# Introduction

The release of ChatGPT in November of 2022 sparked an explosion of interest in generative AI and large language models. The power and flexibility of the new models sparked a surge of creativity as organizations sought to apply new techniques to existing business problems. One of the most popular use cases for generative AI is retrieval-augmented generation (RAG), in which the model accesses additional authoritative information outside its training data set to answer questions and provide information. For example, an organization could provide its health care policy as the external dataset, and then allow the model to field employee questions about policy specifics. RAG offers the ability to customize and refine a general-purpose model for a specific use case without the time and expense of training a new model.

Additionally, the watsonx.governance platform allows organizations to generate explanations of the information provided by the RAG model, which can be critical for verifying the accuracy of the answers the model provides. In the health care policy example, the provided explanations can point to specific portions of the policy used to generate the response, allowing the user to trust that the model output is grounded in factual information, and is not a hallucination.

## About this hands-on lab

In this lab, you will explore this capability by creating a project, associating it with the proper services, and running a Jupyter notebook that will allow you to see this service in action. Finally, you will explore how the Python SDK used in the lab can be easily integrated into an application that allows for the surfacing of the explanation to users.

# Setting up the environment

In this section, you will sign in to the watsonx environment, where you can build, test, deploy, and collaborate on data science and AI projects. Watsonx offers full access management, integration with dozens of databases and data connections, and a full suite of cutting-edge data science and AI tools from prompt labs to no-code rapid prototyping to industry-standard Jupyter notebook runtimes. For this lab, you will create a watsonx project, which organizes data, code, and other assets, and integrates with watsonx.ai and watsonx.governance services.

## Sign into watsonx

1. To begin, you will need to sign in to [watsonx](https://dataplatform.cloud.ibm.com/wx/home?context=wx) using the provided credentials. From your lab reservation screen, scroll down and locate the **Username** and **Password**. Copy and paste these values into a text file.
2. Click on the **IBM Cloud Login** link. The login window opens.

A white background with text

Description automatically generated

1. Enter the username and password into the appropriate fields and click the **Sign in** button. After you are signed in, you will be taken to the **IBM Cloud** dashboard.

A screenshot of a login page

Description automatically generated

1. Click on the hamburger menu in the upper left to open the menu, then click on the **Resource list** menu item. The **Resource list** page opens.

A screenshot of a computer

Description automatically generated

1. Locate the **AI / Machine Learning** item from the list and click on it to expand it. Locate the **Watson OpenScale** item in the expanded section and click on it. Note that the name may differ slightly from the one in the screenshot below.

A screenshot of a computer

Description automatically generated

1. Click on the **Launch watsonx.governance** button. If you are asked to provide any additional information, wait a moment, then click the **Continue** button when it becomes enabled. Note that you may need to wait for up to two minutes for the watsonx.governance services to be prepared.
2. Read the terms and policies if you wish, then check the box to agree, and click the **X** button to close the popup window without taking the tour.

A screenshot of a computer

Description automatically generated

## Generate an API key

To connect to services in your reserved account, you will need an API key, which can be generated via the user interface.

1. Click on the **Manage** button to open the menu, then click on the **Access (IAM)** menu item.

A screenshot of a computer

Description automatically generated

1. If necessary, click on the **Manage identities** menu item to expand it, then click on the **API keys** menu item.

A screenshot of a computer

Description automatically generated

1. Click on the **Create** button. The **Create IBM Cloud API key** window opens.

A screenshot of a computer

Description automatically generated

1. Give your API key a name such as **RAG API key**, then click the **Create** button.

A screenshot of a computer

Description automatically generated

1. Once the key has been created, click the **Copy** button to copy it to your clipboard, then paste the value into the text file that contains your login credentials.

A screenshot of a computer

Description automatically generated

1. Click on the **X** button to close the API key window.

A screenshot of a computer

Description automatically generated

You now have the credentials necessary to proceed with the lab. In the next step, you will create a

## Create a watsonx project

1. Click on the **hamburger menu** in the upper left to expand it, then click on the **Projects** item to expand it. Finally, click on the **View all projects** menu item.

A screenshot of a computer

Description automatically generated

1. Click the **New project** button to create a new project. The **Create a project** screen opens.
2. Fill out the project details, including a **Name** and optional **Description**.
3. Use the **Select storage service** dropdown to select the cloud object storage instance for your environment. If you have multiple options, select the one that most resembles the one in the screenshot below.

A screenshot of a computer

Description automatically generated

1. Click the **Create** button to create your project.

## Associate a machine learning service

For project assets such as Jupyter notebooks to access the IBM Foundation models to perform RAG queries, you must associate a machine learning service with the project. Your lab environment comes with a provisioned machine learning service.

1. From your project screen, click on the **Manage** tab.
2. From the list on the left, click on the **Services & integrations** menu item.
3. Click the blue **Associate service** button. The **Associate service** window opens.

A screenshot of a computer

Description automatically generated

1. Locate and check the box to the left of the Watson Machine Learning service in the list.

A screenshot of a computer

Description automatically generated

1. Click the **Associate** button to associate the service.

## Get the project ID

Getting the ID of the project will allow you to query watsonx.ai Foundation models using the service you just associated.

1. Beneath the **Project** header on the left, click on the **General** menu item.

A screenshot of a computer

Description automatically generated

1. In the **Details** tile, locate the **Project ID** item and click the **Copy to clipboard** icon to copy the value. Paste it into the text file you have been using to store credentials.

A screenshot of a computer

Description automatically generated

Your environment is now fully configured, and you have all of the credentials necessary to access Foundation models. In the next step, you will create a Jupyter notebook to query the models.

# Executing the Jupyter notebook

In this section of the lab, you will load a Jupyter notebook from a GitHub repository, input your credentials, and run it.

## Create the notebook

1. Click on the **Assets** tab of your project.
2. Click on the **New asset** button. The **What do you want to do?** window opens.

A screenshot of a computer

Description automatically generated

1. Scroll down to the **Work with models** section, and click on the **Work with data and models in Python or R notebooks** tile.

A screenshot of a computer

Description automatically generated

1. Click on the **URL** menu option on the right.
2. Give your notebook a name such as **RAG Explanations**.
3. Copy and paste the following value into the **Notebook URL** field:

https://raw.githubusercontent.com/ericmartens/rag-explain/refs/heads/main/source-attribution-using-protodash-for-rag.ipynb

A screenshot of a computer

Description automatically generated

1. Click the **Create** button. The notebook will created in your project, which can take up to two minutes to complete.

## Update the notebook with your API key

Code cell modification here.

1. Step for the sub task.
2. Step for the sub task.

## Execute the notebook

Walk through notebook execution, calling out the output of the cells.

1. Step for the sub task
2. Step for the sub task

# Integrating with an application

You have explored how, using the watsonx Python SDK, you can generate explanations for RAG queries in a notebook. In this section, you will see an example of how that SDK can be used in a simple Flask web application to present both the RAG answers and supporting information in a more visually appealing way, like how an actual organization might present it to its users.

## View the application website

1. Step for the sub task.
2. Step for the sub task.

## Explore the application code

1. Step for the sub task.
2. Step for the sub task.

# Getting help and troubleshooting

This section provides information about getting help with your lab and some common troubleshooting topics.

## Common troubleshooting tips

Use this section to document common troubleshooting issues that users may experience.

## Getting help

Customize this section as needed, keeping in mind that some individuals will be doing this lab in the self-paced room without an instructor and would need guidance on where to go for assistance.