Cloud Pak for Data tutorial: Banking

Learn how to use IBM® Cloud Pak for Data to discover, transform, and analyze data in a private cloud environment.

Learning objectives

You will learn:
- Hosting a Db2® server in a Docker container and loading it with sample data
- Connecting Db2 and Cloud Pak for Data
- Discovering data assets in the enterprise catalog and browsing and searching for data
- Creating, compiling, and running a data transformation job
- Creating an analytics project
- Working with a Jupyter notebook to analyze sample data and train and test a data science model

Time required

90 minutes

- **Introduction**
  In this tutorial you will learn how to use IBM Cloud Pak for Data to discover, transform, and analyze data. While doing so, you'll get a guided tour of this robust end-to-end solution for all of the data and analytic needs within your enterprise.
- **Preparing to use Cloud Pak for Data**
  Before you can use IBM Cloud Pak for Data to generate meaningful insights, you must set up a data source and create a connection to it.
- **Data discovery**
  The data discovery feature in IBM Cloud Pak for Data enables you to catalog data across your enterprise to make it easier to search for, govern, and use the data in different machine learning initiatives. This module focuses on data discovery.
- **Data transformation**
  In this module, you'll learn how to transform some of the data in the BANK database and get the data in better shape.
- **Building a machine learning model for data insight**
  The final module of the tutorial will show you how to use IBM Cloud Pak for Data to unlock the insight in your enterprise data stores.
- **Summary and resources**
  Taking this tutorial gave you a brief introduction to IBM's single, end-to-end platform for data management, data governance, and data science analytics.

Introduction

In this tutorial you will learn how to use IBM® Cloud Pak for Data to discover, transform, and analyze data. While doing so, you'll get a guided tour of this robust end-to-end solution for all of the data and analytic needs within your enterprise. You will download a package that includes a sample database of bank customer and employee records and a preconfigured analysis model that you can use to run a basic machine learning simulation.

As you work with Cloud Pak for Data to find, prepare, and examine the data, you'll gain insight into how you can use the product to collect, govern, and analyze your data.

Learning objectives

In this tutorial you will learn how to:
- Host a Db2® server in a Docker container and load it with sample data
- Create and test a connection between Db2 and Cloud Pak for Data
- Prompt Cloud Pak for Data to discover data assets in its catalog
- Browse and search for data assets
- Create, compile, and run a data transformation project and job
- Create an analytics project
- Work with a Jupyter notebook and connect it to the sample database
- Run the notebook and view the results of your analysis
- Perform a basic test of the machine learning model

Time required

This tutorial should take approximately 90 minutes to finish.

System requirements

- Ensure that you have met the system requirements.
- Install Cloud Pak for DataEnterprise Edition and save the URL of the web client that the installer provides to you.
- Install Git on the machine that you will use for the tutorial.

Prerequisites

- Understanding of database concepts such as a table join
- Basic machine learning knowledge
- Some familiarity with Python is recommended. The Jupyter notebook that is used in the machine learning section runs on Python 3.6.

Preparing to use Cloud Pak for Data

Before you can use IBM® Cloud Pak for Data to generate meaningful insights, you must set up a data source and create a connection to it.

Learning objectives

After completing the lessons in this module you will know how to do the following tasks:

- Host a Db2 server in a Docker container and load it with sample data
- Create and test a connection between Db2 and Cloud Pak for Data

- **Setting up the database and sample tables**
  IBM Cloud Pak for Data supports multiple data source types such as Oracle, Db2®, and HDFS. In this lesson, you'll load and set up data in a Db2 database.
- **Creating and testing a connection**
  In this lesson, you'll create a connection to the Db2 database so that you can use IBM Cloud Pak for Data to discover the assets in the database.

Setting up the database and sample tables

IBM® Cloud Pak for Data supports multiple data source types such as Oracle, Db2®, and HDFS. In this lesson, you'll load and set up data in a Db2 database.

Procedure

1. Log in to the cluster where Cloud Pak for Data is deployed or log in to a Linux-based system (RedHat or Ubuntu) that can access the cluster over your network.
2. From your home directory, clone the tutorial sample files:`git clone`

   `https://github.com/IBM-ICP4D/icp4d-tutorials.git`

   Recommendation: Keep the sample tutorial folders up to date from the GitHub repository so that you can use them for other tutorials to explore the features of Cloud Pak for Data.

3. Change to the tutorials directory:`cd icp4d-tutorials/tutorials/`

   The sample data-loading utility, load_samples.sh, provides an easy way to host a Db2 server and load it with sample data.
4. Run the following command to view the list of sample data that is provided in the load_samples.sh utility:`./load_samples.sh -l`

5. Run the following command to load the sample data into a Db2 database:

   `./load_samples.sh -t banking-001`

   Note: You might experience a period of what seems to be inactivity in your terminal session. Be patient. It takes time to create the Db2 instance and load the data into the database.

6. After the data loading process completes, an instance of Db2 is hosted on your cluster as a Docker container.

Results

Now that you created a hosted instance of the database in your cluster loaded with sample banking data, you'll learn how to connect to this new data source from Cloud Pak for Data.

Creating and testing a connection

In this lesson, you'll create a connection to the Db2® database so that you can use IBM® Cloud Pak for Data to discover the assets in the database.

About this task

You'll enter connection details, such as the name of the host where the Db2 database is located, and the credentials to access the database. **Procedure**

1. Log in to Cloud Pak for Data from a web browser.
2. From the menu, select Connections.
3. On the Connections page, click Add connection.
4. On the Create connection page, enter the connection details for your Db2 database.
   - **Connection name:**`ICP4DTutorial`
   - **Connection type:** Select Db2
   - **Host:** IP address of the server. A mapped host name is not valid for this field.
   - **Port:**`50000`
   - **Database:**`BANK`
   - **Username:**`db2inst1`
   - **Password:**`password`

5. Test the connection by clicking Test connection.When the system confirms that the test was successful and to save your connection settings, click the Add button.

6. On the Connections page, confirm that the table shows your `ICP4DTutorial` data connection.

Results

You successfully created a connection to the Db2 database that is hosted in your Cloud Pak for Data cluster. Next, you'll learn how to discover the assets in the database by using the discovery services.

Data discovery

The data discovery feature in IBM® Cloud Pak for Data enables you to catalog data across your enterprise to make it easier to search for, govern, and use the data in different machine learning initiatives. This module focuses on data discovery.

Learning objectives

After completing the lessons in this module you will know how to do the following tasks:
- Prompt Cloud Pak for Data to discover data assets in the enterprise data catalog
- Browse for data assets
- Use enterprise search to look for data assets

- **Discovering assets**
   In this lesson, you'll direct IBM Cloud Pak for Data to discover the assets in your Db2® database so that you can find them by browsing or searching the enterprise data catalog.
- **Browsing the enterprise data catalog**
   After the data discovery process, you can view and browse the data assets that you added to the IBM Cloud Pak for Data catalog.
- **Using enterprise search**
   In addition to browsing the catalog, you can use the enterprise search feature in IBM Cloud Pak for Data to quickly locate data assets.

Discovering assets

In this lesson, you'll direct IBM® Cloud Pak for Data to discover the assets in your Db2® database so that you can find them by browsing or searching the enterprise data catalog.

About this task

Cloud Pak for Data can search and catalog your entire data source, or you can specify a subset of the data. In this lesson, we will catalog all of the data in the sample data source. **Procedure**

1. From the menu, select Organize > Discover assets.
2. On the Discover assets page, choose the `ICP4DTutorial` connection as the data source.
3. Leave the Schemas field empty so that you can discover all of the assets in the database.To discover and catalog a subset of the assets, you would click Browse and select the tables or schemas to discover.
4. For Select the tasks that you want to run, check all of the following boxes:
   - Profile and classify data
   - Analyze data quality
   - Assign business terms
   - Use data sampling

5. In the Set the maximum number of records that you want to include in your data set sample: field, enter `2000`.
6. Keep the default setting in the Select the method that you want to use to create your sample: list of methods.
7. Click Discover.The discovery job status will first show as `Running` and then show as `Finished` after a refresh.

Results

Next, you'll learn different methods of browsing the discovered assets.

Browsing the enterprise data catalog

After the data discovery process, you can view and browse the data assets that you added to the IBM® Cloud Pak for Data catalog.

About this task

Browsing is a method of data discovery that enables you to navigate through the contents of the data source and choose which assets to view more closely. It is useful when you have an idea of what you are looking for. You can browse asset types or use filters to narrow search results.

Procedure

1. From the menu, select Organize > Data catalog.
2. On the Data exploration page, expand Databases in the Explore Assets panel. Select the following items to reveal what was discovered:
   - Database
   - Database Schema
   - Database Table

   In the All Results panel, the sample database has the following three schemas with multiple tables in each schema:
   - BANK1
   - BANK2
   - BANK3

   .
3. Click the BANK2 schema link in the results panel for our next lesson.Attributes of the database are displayed along with the link to the assets that are contained in this database.

Database Schema Details: 🏛 BANK2

Governance	# BANK2
Contains Assets (3)	IS-EN-CONDUCTOR-0.EN-COND » db2
Lineage Information	

Created by
admin admin

Created on
20 February 2019, 12:00:21 pm

Modified by
admin admin

Modified on
20 February 2019, 12:00:21 pm

Governance

Database
db2

Context
IS-EN-CONDUCTOR-0.EN-COND » db2

---

4. Click the Contains Assets (3) link to view the contents of the BANK2 schema.The schema contains the following three tables:
   - BANK_ACCOUNTS
   - BANK_CUSTOMERS
   - BANK_WRKEX_SALARY

Results

Browsing the data catalog is just one method of discovering the assets in Cloud Pak for Data. You can also use the Enterprise Search feature, which you'll do in the next lesson.

Using enterprise search

In addition to browsing the catalog, you can use the enterprise search feature in IBM® Cloud Pak for Data to quickly locate data assets.

About this task

Using enterprise search might be more appropriate when you are not familiar with your data and are not sure where to find the assets in which you are interested. With this feature, you can search across different data sources for information about a given concept in your enterprise.

The search engine uses text match, asset rating, comments, context match, modification date, quality score, and usage frequency to help you discover the most relevant results.

Procedure

In the Cloud Pak for Data toolbar, click the magnifying glass icon () to reveal the search bar.Searching on `BANK` returns a number of results, three of which you located with the browse feature in the previous lesson:

- BANK_ACCOUNTS
- BANK_CUSTOMERS
- BANK_CLIENTS
- BANK_WRKEX_SALARY

Results

With experience, you'll learn whether search or browsing is the more effective method for the type of data that you want to discover. In our next set of lessons, you'll learn how to use Cloud Pak for Data to transform data to better suit your purposes.

Data transformation

In this module, you'll learn how to transform some of the data in the BANK database and get the data in better shape.

The following screen captures show the columns in the BANK_ACCOUNTS and BANK_CUSTOMERS tables in the database.

BANK_ACCOUNTS	BANK_CUSTOMERS

In the next several lessons, you'll join these two tables on the CUSTOMER_ID column and filter out the EMAIL and PHONE1 columns, which are not needed for the project.

Learning objectives

After completing the lessons in this module, you'll know how to do the following tasks:
- Create a data transformation project
- Create a transformation job
- Add connections to the data flow designer
- Add a join stage to the designer canvas
- Compile and execute the transformation job
- Search for and preview the table that the job generates

- **Creating a data transformation project and job**
   In this lesson, you'll create a data transformation project, called `MyAnalysisProject`, and a transformation job.
- **Adding connections to the canvas**
   The Data Flow Designer is loaded with a palette and a designer canvas. You can choose various connectors and stages with which to build your transformation job.
- **Adding a join stage to the canvas**
   This lesson uses the Data Flow Designer to create a join between the connections to the BANK_ACCOUNTS and BANK_CUSTOMERS tables. Effectively, the join results in a table that contains the data of both the BANK_ACCOUNTS and BANK_CUSTOMERS tables.
- **Compiling and running the job**
   When you compile a job, IBM Cloud Pak for Data validates the design of the job by looking at inputs, transformations, and other details. When the job compiles successfully, it is ready to run.

Creating a data transformation project and job

In this lesson, you'll create a data transformation project, called `MyAnalysisProject`, and a transformation job.

Procedure

1. From the menu, select Organize > Transform data.
2. On the Projects page, create a new project by clicking Create.
3. Name the project `MyAnalysisProject` and click Create.The project console shows the project after it is created. Note: Project creation takes some time. Be patient.

4. To start the job creation process, click the `MyAnalysisProject` project entry in the table.
5. On the Project: MyAnalysisProject page with the Jobs tab selected, click Create. The Data Flow Designer opens.

Results
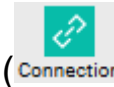
Now that you created a project and started the job creation process, you'll use the Data Flow Designer to build your transformation job.

Adding connections to the canvas

The Data Flow Designer is loaded with a palette and a designer canvas. You can choose various connectors and stages with which to build your transformation job.

Procedure
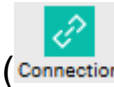
1. In the Connectors palette, click the Connection icon ( ) to select it and then click on the canvas.
2. In the Connection Asset Browser, click Import and select ICP4DTutorial from the list. Click Import.
3. With ICP4DTutorial highlighted, click Next to move to the schema selection page.
4. Select the BANK2 schema and click Next to move to the table selection page.
5. Select the BANK_ACCOUNTS table and click Next to move to the column selection page.
6. With all of the columns selected by default, click Add to Job.The connection to the BANK_ACCOUNTS table is now ready to be used.

7. Follow similar steps to add the BANK_CUSTOMERS table connection to the same canvas. This table is also located in ICP4DTutorial > BANK2.

Results

With the two table connections added to the canvas, you'll add a join stage to the Data Flow Designer canvas.

Adding a join stage to the canvas

This lesson uses the Data Flow Designer to create a join between the connections to the BANK_ACCOUNTS and BANK_CUSTOMERS tables. Effectively, the join results in a table that contains the data of both the BANK_ACCOUNTS and BANK_CUSTOMERS tables.

Procedure

1. In the Stages palette, click the Join icon (Join) and click the canvas to drop it there.The designer canvas now contains icons for BANK_ACCOUNTS, BANK_CUSTOMERS, and Join.

2. Create a link between the first table icon and the Join icon by performing the following steps:
   A. Click on a circular nub at the edge of the BANK_ACCOUNTS icon and then click on a nub at the edge of the Join icon.
   B. A link line appears joining the two icons.
3. Repeat the steps for the BANK_CUSTOMERS icon.With both links added, the canvas looks like the following screen capture:
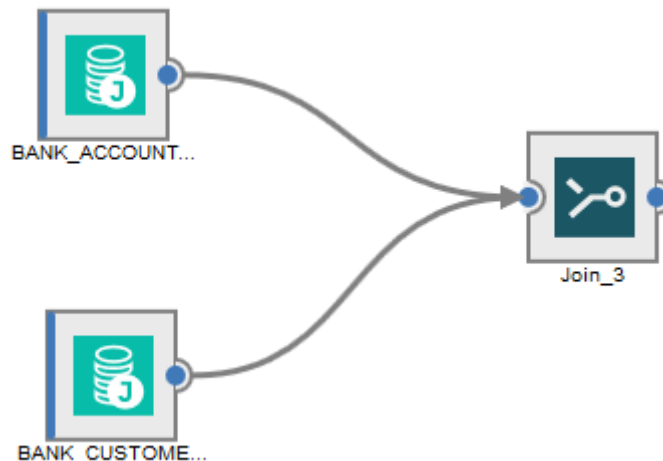
4. Double-click the Join icon to reveal additional properties for the join.On the first tab of the join properties window, you can see that IBM® Cloud Pak for Data has automatically detected that the CUSTOMER_ID column is common between the two tables and recommends an inner join.

5. Click OK to agree to the join type recommendation.
6. To save the result of joining two tables into a single table, return to the Connectors palette and click on the Connection icon. Click on the canvas to drop the icon onto it. In the Connection Asset Browser, add the ICP4DTutorial connection by clicking the Add selected connection as target check box and then click Add to Job.
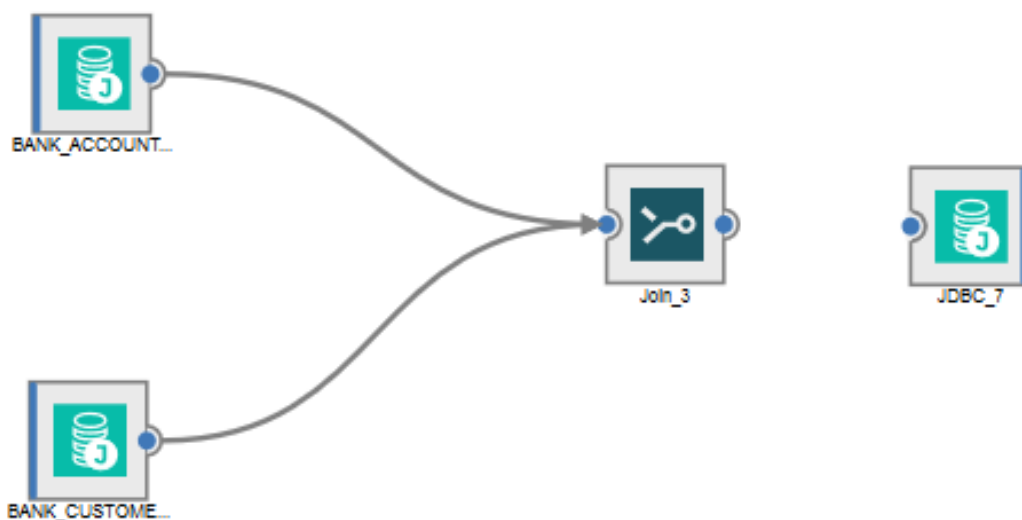
7. Create a link between the first Join icon and the new connection icon by performing the following steps:
   A. Click on a circular nub at the edge of the Join icon and then click on a nub at the edge of the new connection icon.

B. A link line appears joining the two icons.
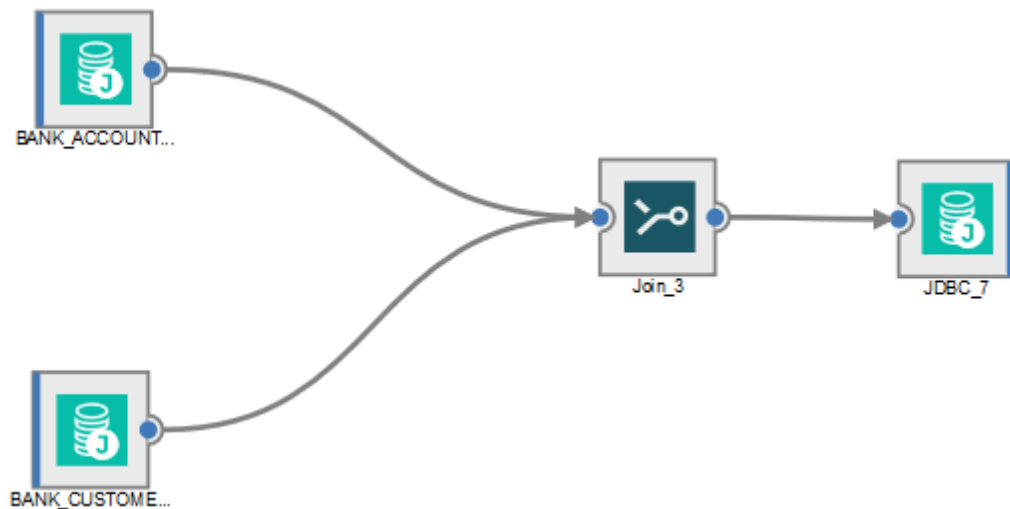
8. After the link is added, double-click the newly created connection icon to open its properties window.
9. In the Table name field, name the generated table ICP4D_CUST_ACC_JOIN. For Table Action, choose Create.

10. Click OK to save your changes.With these steps, the basic transformation job is ready.
11. On the toolbar, click the Save icon.
12. Name the transformation job `Customer_Account_Join` and click Save.

Results

Next, you'll compile the job that you created and then run the job to see the results.

Compiling and running the job

When you compile a job, IBM® Cloud Pak for Data validates the design of the job by looking at inputs, transformations, and other details. When the job compiles successfully, it is ready to run.

About this task

Saving your job in the previous lesson enables the compiling mechanism. In the Job window for your Customer_Account_Join job, the wrench icon is now enabled.
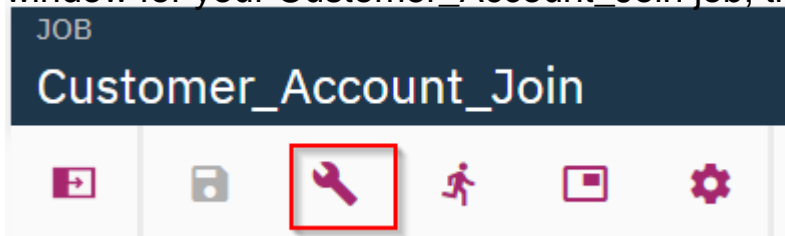
Procedure

1. Click the wrench icon to begin the compiling process.The Job window shows a confirmation message when the process completes.The Run icon ( ) is now enabled in the Job window.

2. Click the Run icon.A Job Run Options window opens, giving you the opportunity to let the job run with the default No limit to warnings or you can select to limit the number of warnings before the job is aborted by clicking in the Warnings field and entering an integer number.


Click Run.

The system will let you know when the job is done. To refresh run status, click the Refresh icon ( ) in the toolbar.When the job finishes, the

ICP4D_CUST_ACC_JOIN table is created in the Db2® BANK database. The table contains the data that was generated by the Customer_Account_Join job.

3. To view the table, click the enterprise search icon (🔍) and search for "ICP4D."In the results, you'll see the new ICP4D_CUST_ACC_JOIN table.
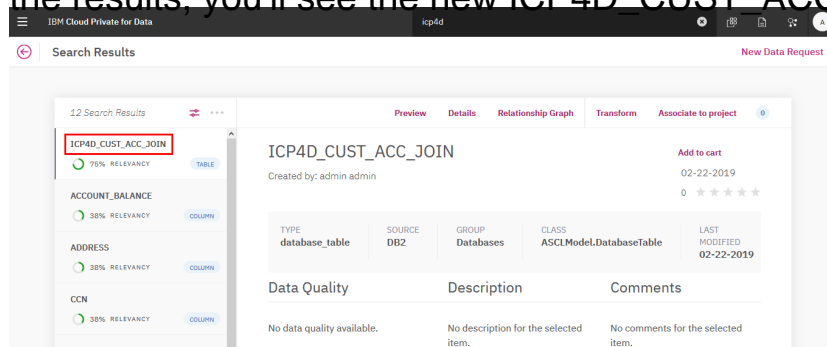
4. To browse the table, click Preview. You will need to enter your user ID and password for the Db2 database (db2inst1/password).The preview window for the ICP4D_CUST_ACC_JOIN table displays.

Preview - ICP4D_CUST_ACC_JOIN

ACCOUNT_ID	CUSTOMER_ID	ACCOUNT_TYPE	ACCOUNT_BALANCE	JOINT_AC
100001	1	SAV	-38735.0	NO
100002	2	SAV	-6282.0	NO
100003	3	SAV	-3475.0	YES

Results

You now know how to compile and run a join job, search for the generated table, and view its contents. In our next set of lessons, you'll learn how to build a simple machine learning model.

Building a machine learning model for data insight

The final module of the tutorial will show you how to use IBM® Cloud Pak for Data to unlock the insight in your enterprise data stores.

You will build a simple regression model to connect to the sample database and predict salaries for new employees who join the bank. The model will be trained using the existing data set within the bank, which is a simplified data set of years of experience and current salaries of employees.

In going through these lessons you will also learn how to use notebooks in data science projects. Notebooks include all of the building blocks you need to work with data: The data, the code computations that process the data, and visualizations of the results.

Learning objectives

After completing the lessons in this module you will know how to do the following tasks:
- Create an analytics project
- Work with a Jupyter notebook and connect it to the sample database
- Run the notebook and view the results of your analysis
- Perform a basic test of the machine learning model

- **Creating an analytics project and working with the project assets**
  In this lesson, you'll learn how to create an analytics project and begin the process of training a simple linear regression model with data.
- **Viewing and testing the analysis model**
  IBM Cloud Pak for Data provides various options for analytics models such as testing, scoring, and publishing to the enterprise catalog. In this lesson, you'll perform a test of the analysis model.

Creating an analytics project and working with the project assets

In this lesson, you'll learn how to create an analytics project and begin the process of training a simple linear regression model with data.

About this task

For the purposes of this tutorial, you'll import an existing project that came with the collateral that you downloaded in Setting up the database and sample tables. The key component is a Jupyter notebook that you'll connect to the Db2 BANK sample database to train the model based on the data set.

Procedure

1. To navigate to the Analytics project console, select Projects from the menu.
2. Create a new project on the Projects page by clicking New project.
3. In the Create a new project window, keep the default selection of Analytics project and enter `ICP4D-Tutorial-Bank001-Analyze` as a name for the project.
4. Because you are importing an existing project on the Create Project page, click the From file tab.
5. On your computer, change directories to the ICP4XTutorial/assets/banking-001 folder, where a project import file from the package that you downloaded from the Git repository is saved.
6. Drag and drop the ICP4D-Tutorial-Bank001-Analyze.zip file to the Project File area on the Create Project page.

Create Project

New   **From file**   From Git repository

Name*

Analytics project

Project File*

Drag and drop your **.zip, .tar, or .gz** file here or
browse your local file system

When the file import is complete, you'll see ICP4D-Tutorial-Bank001-Analyze in the Project File drag-and-drop area.

7. To create the project, click Create.The following project dashboard opens:

The dashboard shows the various project artifacts that were imported.

8. Click Notebooks listed under the Assets tab.The Notebooks area shows that you imported a Predict Salary Jupyter notebook. Assets for analytics projects can

also be other types of notebooks, scripts, and data sets.

A Jupyter notebook is a web-based environment for interactive computing. You can run small pieces of code that process your data, and you can immediately view the results of your computation.

9. From the Predict Salary options menu, select Open in Jupyter with Python 3.6, Spark 2.3.2.The notebook is already annotated with documentation on each section as we build the linear regression model.
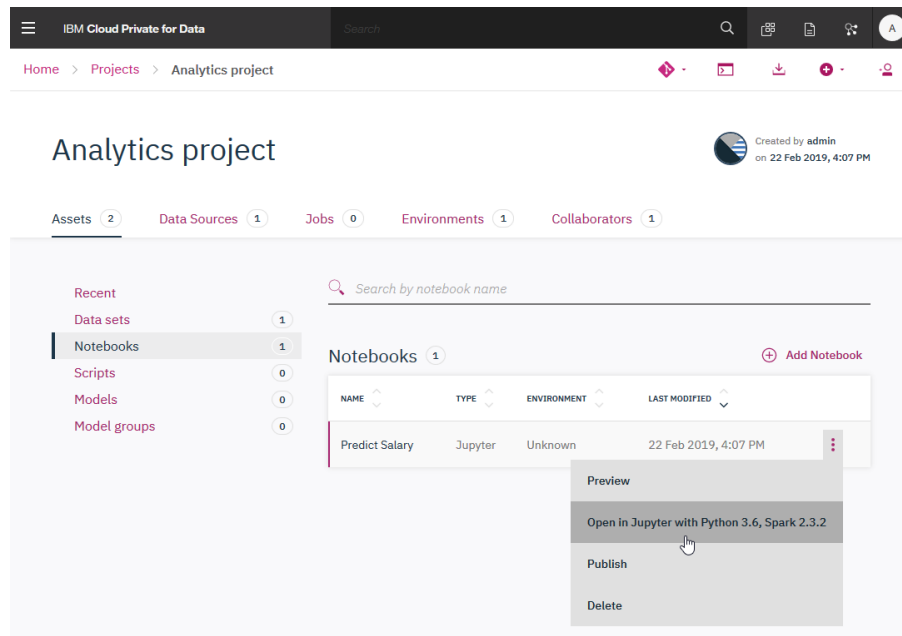
10. Before running the notebook, click the Prepare Environment link and replace `IP_ADDRESS` with the IP address that you entered when you set up the database connection in the Creating and testing a connection lesson of the


tutorial.

This step enables IBM Cloud Pak for Data to connect to the BANK database for the analysis.

In the notebook you can run all of the cells in the notebook or run only the selected cell, which is a great way to explore the model. For this lesson, you'll run all of the cells.
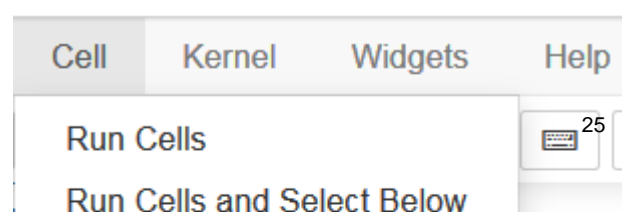
When the notebook is finished running, you should be able to see plots of training and test data against the trained model and see the comparison of actual results versus predicted results.

Salary vs Experience (Training set)

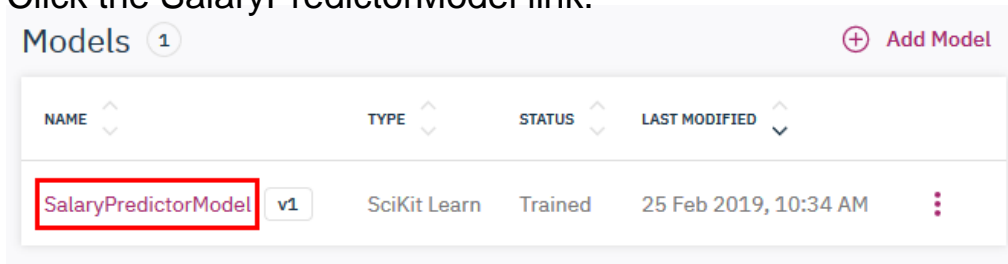Salary vs Experience (Test set)

Predicted and Actuals
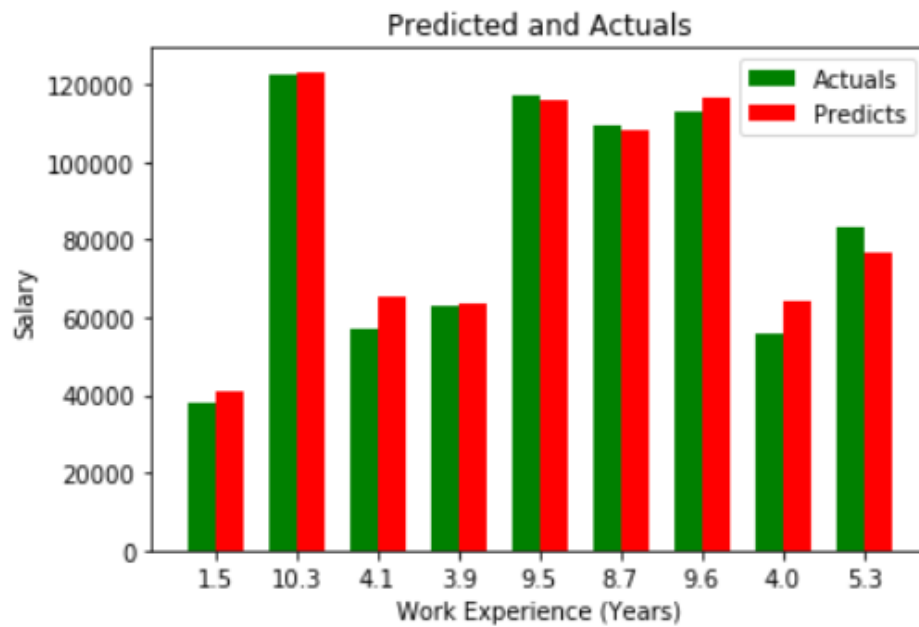
Results

After learning how to work with a Jupyter notebook, in the next lesson you'll see how to enter data to test the regression model.

Viewing and testing the analysis model

IBM® Cloud Pak for Data provides various options for analytics models such as testing, scoring, and publishing to the enterprise catalog. In this lesson, you'll perform a test of the analysis model.

Procedure

1. At the top of the Predict Salary notebook page, click the ICP4D-Tutorial-Bank001-Analyze project name to return to the project assets view.
2. Click Models listed under the Assets tab.The Models area shows that you imported the SalaryPredictorModel machine learning model.
   Let's test the machine learning model with some data.

3. Click the SalaryPredictorModel link.

The SalaryPredictorModel view shows that this is a Linear Regression Model built with the Python 3.6 engine.

4. Select the Real-time score tab and enter the number 4 in the EXPERIENCE_YEARS field. Click Submit.The model analysis yields a predicted salary of $64,218.44 for an employee with four years of experience.

Results

While this was a simple demonstration, it nevertheless gives you a glimpse into the kind of data science analysis that you can do with Cloud Pak for Data. You can also create Scala or Zeppelin notebooks that run on Spark, or you can set up and start using RStudio. To create models, you can use APIs from notebooks, import models, or use the model builder to create more complex projects.

Summary and resources

Taking this tutorial gave you a brief introduction to IBM's single, end-to-end platform for data management, data governance, and data science analytics.
IBM® Cloud Pak for Data provides a powerful means for data scientists, data engineers, and data stewards to collaborate in acquiring, governing, and extracting insights from their enterprises.

Lessons learned

By completing this tutorial, you learned about the following concepts and tasks:
- Using Docker with Cloud Pak for Data to host a data source and connect it to the platform
- Discovering data assets in the enterprise catalog through search and browsing
- Transforming data to improve its usefulness in analytics projects
- Working with data science notebooks to run and test machine learning models

Additional resources

- Data shaping and cleansing
- Data science
- Artificial intelligence
- Python
- Data analytics code patterns
- AI and data code pattern playlist
- Master the art of data science with IBM's Watson Studio