

Lab Center – Hands-On Lab

Session 5934

Trust the Art of Insight in the Age of AI through the
Use of Data Governance

sanjitc@us.ibm.com



DISCLAIMER

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results like those stated here.

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.

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 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.

It is the customer's responsibility to insure 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.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. 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: www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates. Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenShift is a trademark of Red Hat, Inc.

UNIX is a registered trademark of The Open Group in the United States and other countries.

© 2020 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.

We Value Your Feedback

Don't forget to submit your Think 2020 session and speaker feedback! Your feedback is very important to us – we use it to continually improve the conference.

Access the Think 2020 agenda tool to quickly submit surveys from your smartphone or laptop.

Table of Contents

1 Introduction	5
2 Prerequisites	6
3 Getting Started	7
4 Access Credentials	8
4.1 Sign into Cloud Pak for Data web console as Administrator	8
4.2 Sign into Cloud Pak for Data web console as student<id>	11
5 Create Global Connection	12
6 Create a Knowledge Catalog.....	14
7 Create Analytic Project	16
7.1 Create the Travel Analytic Project	16
7.2 Create the Auto Discovery Analytic Project	18
8 Discover and Catalog Data Assets	19
8.1 Download Documents Catalog	19
8.2 Add Unstructured Data to Catalog	19
8.3 Auto Discover Data Assets.....	24
8.4 Review Discovery Result	27
8.5 Add Structured Data to Catalog.....	31
9 Discover Assets using Quick Scan	34
10 Implement Business Glossary	38
10.1 Download Business Glossaries.....	38
10.2 Import categories	38
10.3 Import Terms	40
10.4 Automated Discovery	41
11 Data Assets Sentiments.....	43
12 Shop for Data.....	47
12.1 Shop Data Assets using Suggestion.....	47
12.2 Shop Data Assets using Search	48
13 Prepare Data for Analytics Project.....	49
13.1 Add Catalogued Data Assets to Analytics Project	49
13.2 Refine Data	51
13.3 Combine Data	53
13.4 Rename Column	56
13.5 Create a New Column.....	57
13.6 Run Data Flow.....	58
14 Congratulation!	60

1 Introduction

AI enabling assets such as ML models within the enterprise are the next generation of information assets in today's world. Cloud Pak for Data with governance can enable a culture where data governance is an enabler of workforce productivity through self-service as well as a conformance force for regulatory obligations. How does one extend the concepts of data Governance to AI? Is it even possible? Cloud Pak for Data from IBM can make this a reality. This lab shows how the use of foundational Governance services to make this a natural evolution of Enterprise Data Governance.

2 Prerequisites

- Access to an operational Cloud Pak for Data (v.2.5 or higher) cluster
- Watson Knowledge Catalog (WKC v.3.0) deployed on the CPD cluster

3 Getting Started

You just need Cloud Pak for Data (CPD), User Interface (UI) for this lab. Each student needs to use their unique URL to access the CPD UI. URL can be found with the Load Balancer entry under Ports column. For example: <https://services-uscentral.skytap.com:xxxxx>

Use the Google Chrome to access the CPD UI.

Lab Guide		Assigned					Ports	
Ident	Assigned To	Student	Username	Password	Links			
S0001	1				Skytap Console	Skytap View-only	<ul style="list-style-type: none"> - B - Master 1:22 (10.1.1.2) - services-uscentral.skytap.com:18710 - A - Load Balancer:31419 (10.1.1.1) - services-uscentral.skytap.com:18745 - C - Master 2:22 (10.1.1.3) - services-uscentral.skytap.com:18746 - D - Master 3:22 (10.1.1.4) - services-uscentral.skytap.com:18747 	

By default, all VMs should already started and shows as running with green annotation. If VMs not started, Go to the **Skytap Console** and click **start** icon on top right corner to start VMs.

VMs: 11		Actions					Sort by name	
<input checked="" type="checkbox"/>								
	Running							
<input checked="" type="checkbox"/>	A - Load Balancer Endpoints: 1 (host-1 - 10.1.1.1)							
<input checked="" type="checkbox"/>	B - Master 1 Endpoints: 1 (rhelhost1 - 10.1.1.2)							
<input checked="" type="checkbox"/>	C - Master 2 Endpoints: 1 (host-2 - 10.1.1.3)							
<input checked="" type="checkbox"/>	D - Master 3 Endpoints: 1 (host-3 - 10.1.1.4)							

4 Access Credentials

4.1 Sign into Cloud Pak for Data web console as Administrator

You should have an operational Cloud Pak for Data Instance. Use latest version of Firefox or Google Chrome browser to access the Cloud Pak for Data UI. Starting from here all instruction need to execute on Cloud Pak for Data UI only.

Login as **admin** to give necessary privileges to the individual user **student<id>** that you are going to use later on this lab.

WELCOME TO
IBM Cloud Pak for Data

Sign in Sign up



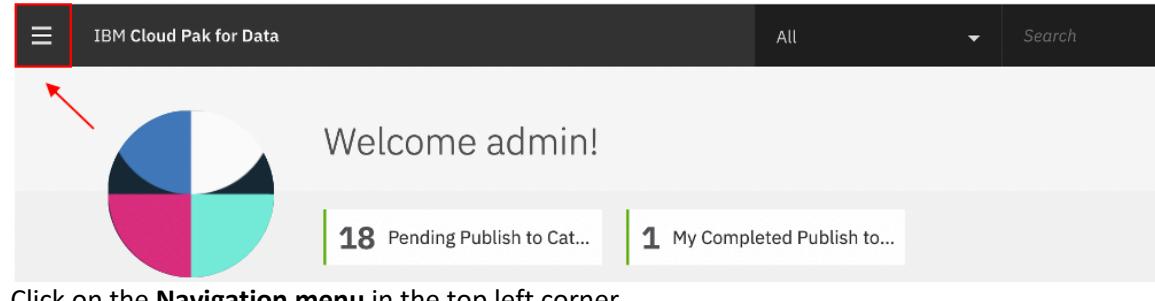
Username
admin

Password

Sign in

A screenshot of the IBM Cloud Pak for Data sign-in page. It features a logo icon of a computer monitor with a key and a user icon. The sign-in form has fields for 'Username' (containing 'admin') and 'Password' (containing '*****'). A red arrow points to the 'Sign in' button at the bottom right.

- Enter username **admin**
- Password is **password**
- Click the **Sign In** button



Click on the **Navigation menu** in the top left corner.

IBM Cloud Pak for Data

Filter navigation

- Home
- Projects
- Connections
- My instances
- Collect
- Analyze
- Administer
 - Manage platform
 - Configure platform
 - Gather diagnostics
 - Manage users**

Click **Manage users**

IBM Cloud Pak for Data

All

Search

Administrator > Manage users

Manage users

Configure LD/

Name	Status	Username	Date added	User ID	Roles
admin	✓	admin	--	1000330999	Administrator, Data Scientist + 4 more
student1	✓	student1	7 Jan, 2020 2:25 PM	1000331001	Business Analyst, Data Engineer + 4 more
student10	✓	student10	7 Jan, 2020 2:30 PM	1000331010	Business Analyst, Data Engineer + 4 more
student11	✓	student11	7 Jan, 2020 2:30 PM	1000331011	Business Analyst, Data Engineer + 4 more

New user

Edit your individual user **student<id>** by using the pen icon on the right side

IBM Cloud Pak for Data

All

Search

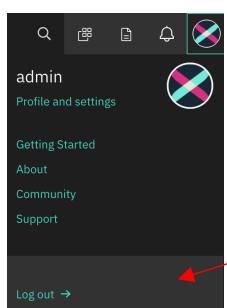
Administrator > Manage users > Users > Edit user

Edit user

User *	student1	Roles *
		<input checked="" type="checkbox"/> Administrator
		<input checked="" type="checkbox"/> Business Analyst
		<input checked="" type="checkbox"/> Data Engineer
		<input checked="" type="checkbox"/> Data Quality Analyst
		<input checked="" type="checkbox"/> Data Scientist
		<input checked="" type="checkbox"/> Data Steward
		<input checked="" type="checkbox"/> Developer

Cancel Save

- Grant the **Administrator** role to your individual user **student<id>**
- Click **Save**



Log out from **admin** user

4.2 Sign into Cloud Pak for Data web console as student<id>

Login as your individual student<id> to continue with remaining lab.

WELCOME TO
IBM Cloud Pak for Data

Sign in Sign up



Username
student1

Password
.....

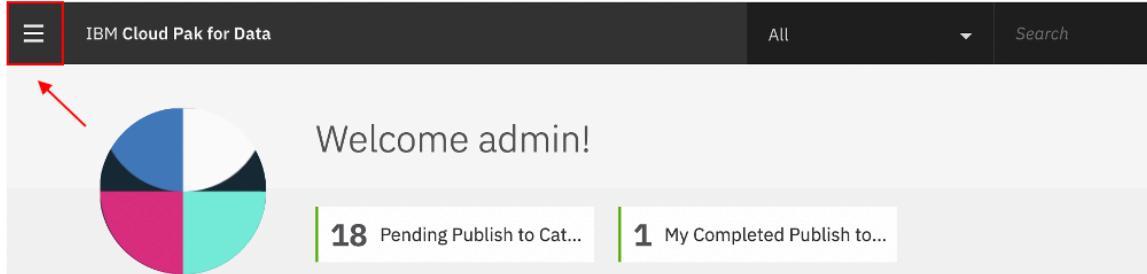
Sign in

- Enter username **student<id>** (Please use your unique student id)
- Password is **password**
- Click the **Sign In** button

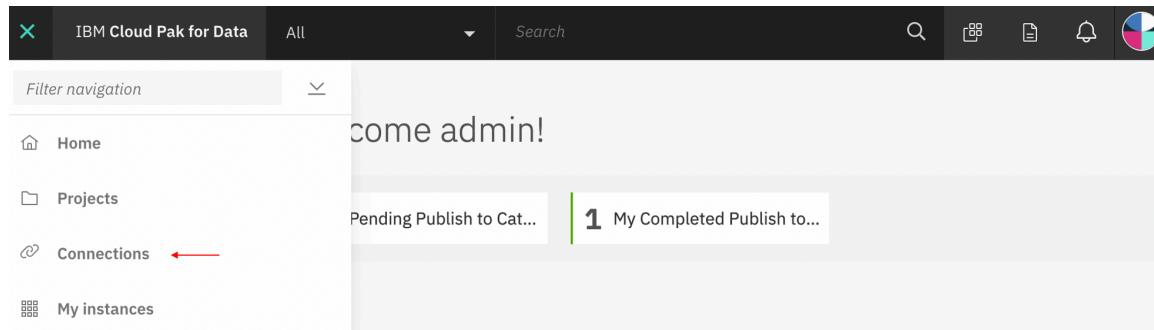
5 Create Global Connection

You need a connection asset for a data source in an analytics project, so you can access data to or from it. A global connection contains necessary information to access respective data source.

In this task you create a global connection to access necessary data from a Db2 database.



Click on the **Navigation menu** in the top left corner.



- Click **Connections**
- Next click **Add connection**

Add connection

Connection name *

Db2 Travel

Description

Connect to Db2 Travel database for Knowledge Catalog

JDBC URL ⓘ

jdbc:db2://169.62.88.219:50000/TRAVEL

Options ⓘ

Type additional options here

Use SSL

Connection type *

Db2

Host *

50000

Port *

50000

Database *

TRAVEL

Username *

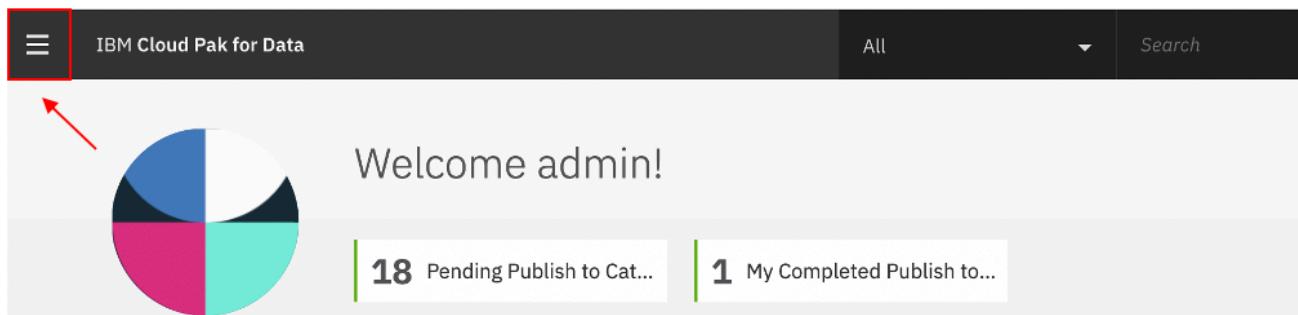
Cancel Test connection Add

- Enter a Connection Name of **Db2 Travel**.
- Enter a Description of **Connect to Db2 Travel database for Knowledge Catalog**.
- Connection type of **Db2**.
- Use the **169.62.88.219** in the Hostname or IP Address field.
- Port of **50000**
- Enter a Database of **TRAVEL**.
- Enter a Username of **db2inst1**.
- Use **password** for the Password field.
- Click **Test connection** for verify the connection.
- Click the **Add** button.

Note: If you are sharing this cluster with other students, connection could be already created by others and you can just use it.

6 Create a Knowledge Catalog

In this task you create a governed Knowledge Catalog. Watson Knowledge Catalog provides a secure enterprise catalog management platform that is supported by a data governance framework. A catalog connects data and knowledge with the people who need to use it. The data governance framework ensures that data access and data quality are compliant with your business rules and standards.



Click on the **Navigation menu** in the top left corner.

A screenshot of the IBM Cloud Pak for Data interface with the navigation menu open. The left sidebar shows various options: Home, Projects, Connections, My instances, Collect, Organize, All catalogs (which has a red arrow pointing to it), Information assets, and Data and AI governance. The main content area displays a welcome message "come admin!" and two status indicators: "Pending Publish to Cat..." and "1 My Completed Publish to...". A large progress bar at the bottom indicates "63%".

Click **Organize > All catalogs** to access catalogs.

Your catalogs

Admin
Enterprise Default Catalog

Creator: admin Date created: Mar 18, 2020 8:27 PM Default catalog configured for sync

New Catalog

Click the **New Catalog** button in the top right corner.

New catalog

Details

Name*
Travel

Description
Travel Knowledge Catalog

Permanently enforce data protection rules
 Enforce data protection rules
 ⓘ Can't be undone. You can't disable protection rule enforcement for a catalog after you enable it.

Permanently enforce data protection rules ?
 Are you sure you want to enforce data protection rules for this catalog? You cannot disable protection rules enforcement after you create the catalog.

Cancel OK

Create

- Enter a Name of **Travel<id>**.
- Enter a Description of **Travel Knowledge Catalog**.
- Select the **Enforce data policies** checkbox.

The **Permanently enforce data protection rules** warning dialog will be displayed, asking if you are sure you want to set this option and informing you that the setting is permanent.

- Click the **OK** button.

By default, access to data assets in a catalog is only restricted by the privacy settings of the data assets. Privacy settings and policy rules can limit which members of the catalog can view and use the assets. You can implement data policies to restrict access to data based on the contents of the data. Data policies help you control data access and ensure that the right people can access the right data. Selecting the option to **Enforce data protection rules** enables the enforcement of data policy rules to allow or deny access to a data asset or redact, substitute or mask data at the column level.

Setting this option for a catalog is a good best practice. Once it is enabled, it cannot be undone, but it does not restrict or impede any functionality, it provides additional security measures to protect data assets.

- Click the **Create** button.

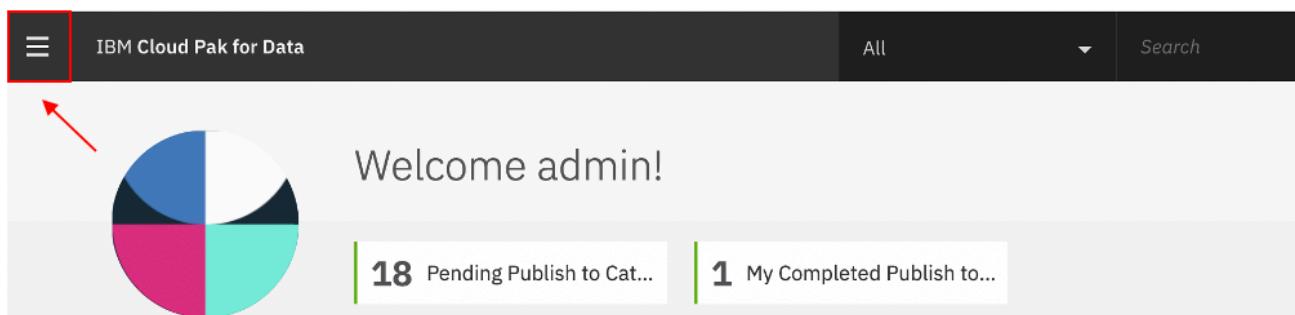
7 Create Analytic Project

In this task you will create two analytical projects.

The first project, which will be named **Travel**, will be used for developing a R application, to collaborate and build the analytic and AI assets et. You will add travel related data assets from the **Travel** knowledge catalog to this project and do some shaping of the data using the **Data Refinery** to prepare the data for analytical insights.

The 2nd project, which will be named **Auto Discovery**, will be used to demonstrate the auto discovery capabilities of Watson Knowledge Catalog. When you catalog a **Connection**, you can choose the option to automatically discover data assets. The discovered assets are added to a analytical project as a temporary holding area for review. You will use a separate project for auto discovery, so it does not disrupt the data analytics project. Once data assets are discovered and added to a project, you can review them, determine which assets are relevant and then publish them to a Knowledge Catalog.

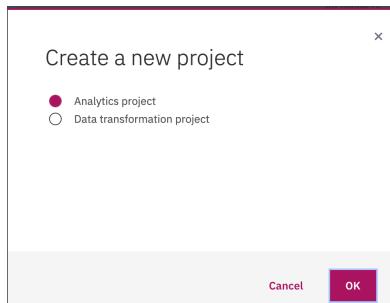
7.1 Create the Travel Analytic Project



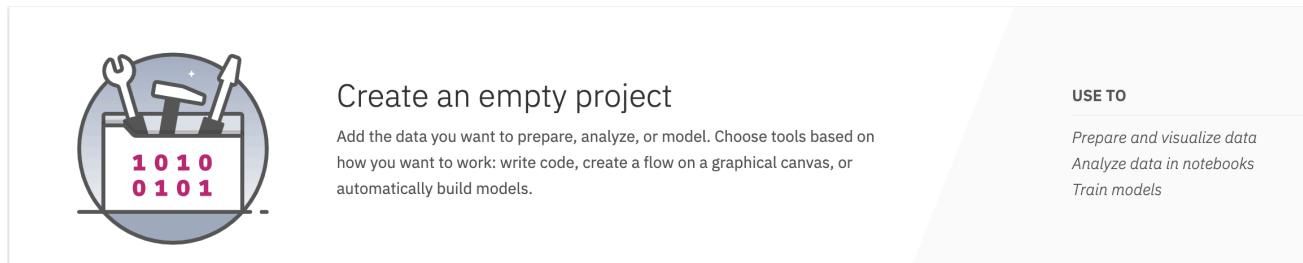
Click on the **Navigation menu** in the top left corner.

Click Projects

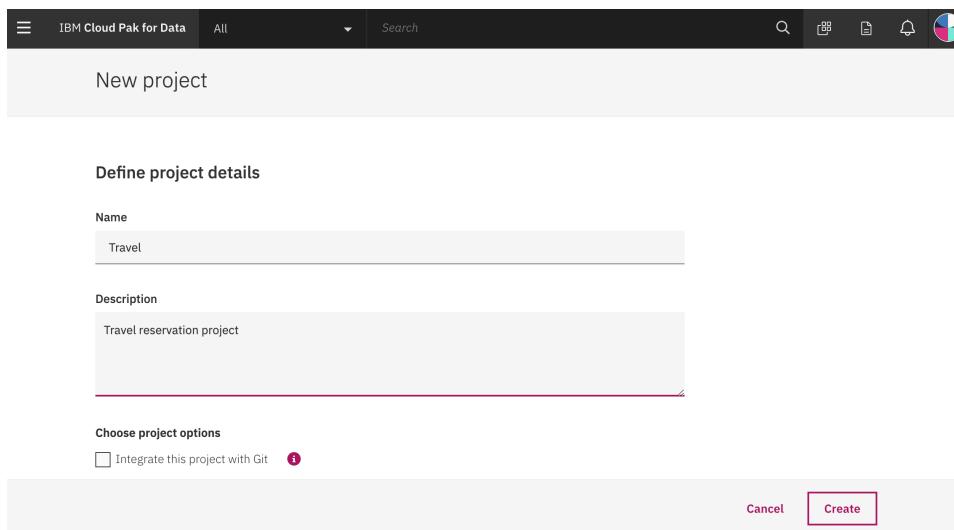
Click New project



If Create a new project popup window appeared, Choose **Analytics project** and click **OK**.



Click on **Create an empty project**.



- Enter a Name of **Travel<id>**.
- Enter a Description of **Travel reservation project**.
- Click the **Create** button.

7.2 Create the Auto Discovery Analytic Project

Same as the **Travel** project, create second analytical project called **Auto Discovery**.

- Enter a Name of **Auto Discovery<id>**.

The screenshot shows the 'IBM Cloud Pak for Data' interface. At the top, there is a navigation bar with 'IBM Cloud Pak for Data' and 'All' selected. A search bar is present, along with several icons: a magnifying glass, a document, a bell, and a profile icon. Below the navigation bar, a button labeled 'New project' is visible. The main area is titled 'Define project details'. It contains two input fields: 'Name' with 'Auto Discovery' typed in, and 'Description' with 'Auto Discovery project'. Under 'Choose project options', there is a checkbox labeled 'Integrate this project with Git' with a small 'i' icon next to it. At the bottom right of the form, there are 'Cancel' and 'Create' buttons, with 'Create' being highlighted.

- Enter a Description of **Auto Discovery project**.
- Click the **Create** button.

8 Discover and Catalog Data Assets

In this task, you will discover and catalog unstructured data assets from the local file system and structured data assets from an existing **Db2 database**. This will introduce you to the three methods available to discover and catalog data assets; **Local files**, **Connected asset** and **Connection**. You will use these methods to catalog data assets into the newly created Knowledge Catalog and then tag them for users to easily find them, understand their content and make them available throughout Cloud Pak for Data, for use during data preparation and within models, dashboards and notebooks.

8.1 Download Documents Catalog

First download two sample document from GIT in your local machine.

- Go to <https://github.com/sanjitc/TravelBid/tree/master/assets>

The screenshot shows a GitHub repository page for 'sanjitz / TravelBid'. The repository has 1 star and 0 forks. It shows a list of files in the 'assets' directory. The files listed are:

- BusinessGlossary (Creating travel-business-glossary-term.csv file, 22 days ago)
- Building_Travel_Reservation_Application_wi... (Added document Think2020 Lab, 12 minutes ago)
- Knowledge-Graph-With-WKC.docx (Added document Think2020 Lab, 11 minutes ago)

- Download **Building_Travel_Reservation_Application_CPD.pdf** and the **Knowledge-Graph-With_WKC** files. Choose on each individual file and click **Download** tab to start download.

8.2 Add Unstructured Data to Catalog

You can see a preview of the contents of many types of data assets and some types of analytical assets in projects and catalogs. These unstructured data can be PDF, text, images, Microsoft Excel documents etc.

You must have Editor or Admin role in a catalog to add a data asset from a file

Click on the **Navigation menu** in the top left corner and access catalog from **Organize > All catalogs**.

The screenshot shows the 'Your catalogs' section in the navigation menu. A catalog named 'Travel' is selected. A context menu is open over the catalog entry, showing options:

- New Catalog (disabled)
- View (selected and highlighted in blue)
- Delete

- Click on the configuration icon of the **Travel<id>** catalog.
- Select **View**

The screenshot shows the IBM Cloud Pak for Data Catalog interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and 'All'. Below it, a sub-navigation bar shows 'Catalogs > Travel'. On the right side, there's a toolbar with icons for search, file operations, and notifications. A dropdown menu is open under the 'Add to Catalog' icon, with options: '+ Add to Catalog', 'Local files', 'Connected asset', and 'Connection'. A red arrow points to the '+ Add to Catalog' option. Below the toolbar, a message says 'Now you can add assets!' with a link to get started.

- To add a local file to catalog, click on **Add to Catalog > Local files**.

The screenshot shows a 'Select File(s)' dialog box. It has a title 'Select File(s)' and instructions: 'Drop your files here or browse your files to add new files (up to 5 GB each). Files for assets are saved in the catalog's associated storage.' Below this is a 'Target' section with a warning: '⚠ Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' At the bottom are 'Cancel' and 'Add' buttons.

- Click the **browse** link in the **Select File(s)** section to bring up the file selection dialog.

The screenshot shows an OS X Finder window titled 'HOL_Assets'. The sidebar on the left lists 'Favorites' (Recents, Applications, Desktop, Documents, Downloads), 'Locations' (Remote Disc), 'Media' (Music, Photos, Movies), and 'Tags'. The main pane displays a list of files: 'Knowledge-Graph-With-WKC' (Word document, 830 KB) and 'Building_Travel_Reservation_Application_with_CPD.pdf' (PDF Document, 2 MB). The file 'Building_Travel_Reservation_Application_with_CPD.pdf' is selected. At the bottom are 'Options', 'Cancel', and 'Open' buttons.

- Locate the “**Building_Travel_Reservation_Application_CPD.pdf**” and the “**Knowledge-Graph-With_WKC**” files under the “**Downloads**” directory on your local machine. Select both using the **Ctrl or Command key** on your keyboard (Ctrl+Click for Windows and Command+Click for MacOS).
- Click the **Open** button to begin cataloging the files.

Catalogs > Travel

Add data assets from local files

Selected Files (2)*

Continue adding files in the drop zone below or [browse](#) to select files

Asset Name	Format	
Knowledge-Ggraph-With-WKC.docx	Microsoft Word	(X)
Building_Travel_Reservation_Application_with_CPD.pdf	PDF	(X)

[Edit name and format](#)

- Click the **pencil icon** next to the **Edit name and format** link.

This allows you to rename the data assets and change their file format. A default file format is inferred for you based on the file extension. In this case, they are PDF and DOCX files, so **PDF** and **Microsoft Word** were auto selected. You **will not** change the format, but you will change their names by removing the file extension.

Catalogs > Travel

Add data assets from local files

Selected Files (2)*

Asset Name	Format
Knowledge-Ggraph-With-WKC	application/vnd.openxmlformats-officedocument.wordprocessingml.document
Building_Travel_Reservation_Application_with_CPD	application/pdf

[Cancel](#) **Apply**

- Click in the **Asset Name** area of the “**Building_Travel_Reservation_Application_CPD.pdf**” file. Go to the end of the filename and remove the **.pdf** extension.
- Click in the **Asset Name** area of the “**Knowledge-Graph-With_WKC.docx**” file. Go to the end of the filename and remove the **.docx** extension.
- Click the **Apply** button to save the filename changes.

The screenshot shows the 'Catalogs' section of the IBM Cloud Pak for Data interface. A file named 'Knowledge-GGraph-With-WKC' in Microsoft Word format and another named 'Building_Travel_Reservation_Application...' in PDF format have been selected. The 'Tags' field contains 'Travel'. The 'Description' field contains 'Travel related document'. On the right side, there are sections for 'Business Terms', 'Tags' (with 'Travel' selected), 'Classification' (set to 'None'), 'Privacy' (set to 'Public'), and 'Members'. A note at the bottom left says '⚠ Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' A 'Cancel' button and a highlighted 'Add' button are at the bottom right.

- Enter a Description of **Travel related document**.
- Enter a Tag of **Travel** into the Tags area.
- Click the + sign next to the tag to add it.

Every time you enter a tag, you need to select the + sign to add the tag. The tags will appear as added tags in the tag area below the tag name. Once a tag is added, it can be used and selected for other data assets. Knowledge Catalog displays all available tags once they are added to the catalog. You will see this in action when you add the next file to the catalog.

The screenshot shows the 'Catalogs' section of the IBM Cloud Pak for Data interface. A file named 'Knowledge-GGraph-With-WKC' in Microsoft Word format and another named 'Building_Travel_Reservation_Application...' in PDF format have been selected. The 'Tags' field contains 'Document'. The 'Description' field contains 'Travel related document'. On the right side, there are sections for 'Business Terms', 'Tags' (with 'Document' selected), 'Classification' (set to 'None'), 'Privacy' (set to 'Public'), and 'Members'. A note at the bottom left says '⚠ Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' A 'Cancel' button and a highlighted 'Add' button are at the bottom right.

- Enter a Tag of **Document** into the Tags area.
- Click the + sign next to the tag to add it.

The screenshot shows the 'Travel' asset being cataloged. The 'Selected Files (2)*' section lists two files: 'Knowledge-Ggraph-With-WKC' (Microsoft Word) and 'Building_Travel_Reservation_Applicatio...' (PDF). The 'Target' section is set to 'Travel'. The 'Description' section contains the text 'Travel related document'. On the right side, there are sections for 'Business Terms', 'Tags', 'Classification', 'Privacy', and 'Members'. A note at the bottom left says '⚠ Stay in the catalog until loading is complete! If you leave the catalog, the incomplete asset will be deleted.' A 'Cancel' button and a highlighted 'Add' button are at the bottom right.

The screenshot shows the **Travel** and **Document** tags that you should have entered for this asset. Make sure you have added them before you proceed to the next step that catalogs them.

- Click the **Add** button to catalog the unstructured data assets.

The screenshot shows the results of a search for 'Travel'. The 'Any type' and 'Any source' filters are applied. The search results show two items: 'Knowledge-Ggraph-With-WKC' and 'Building_Travel_Reservation_Applicatio...'. Both items are listed as 'Data asset'. The 'Tags' column for both items shows 'Travel' and 'Document'. The 'Owner' column shows 'admin' and the 'Added' column shows 'Apr 04, 2020 7:29 PM..PM'. Below the search results, a table shows the details of the cataloged assets.

	Name	Owner	Tags	Business Terms	Type	Date Added
<input type="checkbox"/>	Building_Travel_Reservation_Applicatio...	A admin	Travel Document		Data asset	Apr 04, 2020
<input type="checkbox"/>	Knowledge-Ggraph-With-WKC	A admin	Travel Document		Data asset	Apr 04, 2020

Upon completion, the data assets will automatically be added to the **Browse Assets** section of the catalog asset browser. Scroll down and you will see the two newly added documents in the catalog with the tags you specified. Notice that the **Travel** and **Document** tags have been added to the **Tags** filter area.

8.3 Auto Discover Data Assets

Go to the **Navigation menu** in the top left corner and click on **Travel** catalog from **Organize > All catalogs**.

Your catalogs

New Catalog

Admin

Travel

Creator: admin Date created: Apr 04, 2020 2:11 PM

Travel Knowledge Catalog

⋮

View

Delete

- Click on the **configuration** icon of the **Travel** catalog.
- Select **View**

IBM Cloud Pak for Data

All

Search

Catalogs > Travel

+ Add to Catalog

Browse Assets Access Control Settings

Local files

Connected asset

Connection

- Click **Add to Catalog > Connection** from the catalog menu.

IBM Cloud Pak for Data

All

Search

New connection

Create new Create from global connections (10)

hive-connection Apache Hive

BANK DB Db2

mortgage-informix Informix

hive-hdp3 Apache Hive

Data Virtualization Db2

Oracle Oracle

Db2 Travel Db2

mortgage-db2 Db2

db2_mortgage Db2

new_informix Informix

- Choose **Create from global connection** tab
- Click on the **Db2 Travel** connector.

New connection (Db2 Travel - Db2)

Name *

Description
Connect to Db2 Travel database for Knowledge Catalog

Database *

Host name or IP Address *

Port *

SSL Certificate

Connection discovery Discover data assets

Project for discovered assets *

Credentials
 Personal Shared

All project collaborators use the provided credentials for the connection.

User name * Password *

Enter information for the selected data source

Screenshot

Most of connection information is pre-populated based on the global connection

- Click the **Discover data assets** check box under Connection discovery.
- Select the **Auto Discovery<id>** project from the “Project for discovered assets” selection list.
- Choose **Shared** for Credentials
- Click **Test**
- Click the **Create** button.

Add a tag to the connection you created

Catalogs > Travel

Watson Recommends Highly Rated Recently Added

Name	Owner	Type	Date Added
Building_Travel_Reservation_Application...	admin	Data asset	Apr 04, 2020 7:29 PM,PM
Knowledge-GGraph-With-WKC	admin	Data asset	Apr 04, 2020 7:29 PM,PM
Db2 Travel	admin	Connection	Apr 06, 2020 3:54 PM,PM

- Click on the connection name **Db2 Travel**.

The screenshot shows the IBM Cloud Pak for Data interface. In the top navigation bar, 'IBM Cloud Pak for Data' is selected. Below it, the breadcrumb path shows 'Catalogs > Travel > Db2 Travel'. On the right, there are icons for search, refresh, download, and add to catalog. The main content area is titled 'Db2 Travel' under 'CONNECTION'. It includes sections for 'Description' (connecting to Db2 Travel database for Knowledge Catalog), 'Business Terms' (no terms available), 'Tags' (highlighted with a red arrow), 'Reviews' (0 reviews), and 'Connection' (Source: Db2). At the bottom right, there are 'Remove', 'Download', and 'Add to Project' buttons.

- Hover next to the **Tags** section and Click the **pencil icon** to add tags to the connection.

The screenshot shows the same interface as above, but the 'Tags' section is now being edited. A red arrow points to the 'Start typing to add values' input field. Two tags, 'Db2' and 'Travel', are listed below it. The 'Apply' button is highlighted with a red border. The rest of the interface remains the same, including the 'CONNECTION' section and other details.

- Click in the **Tags** section and select the **Travel** tag from the list of tags and Click the **+** sign next to the tag to add it.
- Similar way adds another tag **Db2**. But you may need to type the tag name.
- Click the **Apply** button.

The discovery process has been running as a service in the background, discovering and populating data assets into the **Auto Discovery** project. Let's examine the project to review the discovery results. We are interested in finding relevant travel data, specifically Rentalcar and Hotel data that will be used by the R application development.

8.4 Review Discovery Result

Go to the **Navigation menu** in the top left corner and click on **Projects**. Select **Auto Discovery<id>** from list of projects.

Overview Assets Environments Jobs Access Control Settings

- Click **Add to project > Connection**

Create new Create from global connections (10)

hive-connection Apache Hive	hive-hdp3 Apache Hive	Db2 Travel Db2	db2_mortgage Db2
BANK DB Db2	Data Virtualization Db2	mortgage-db2 Db2	new_informix Informix
mortgage-informix Informix	Oracle Oracle		

- Choose **Create from global connection tab**
- Click on the **Db2 Travel** connector.

Name *

Description
 Connect to Db2 Travel database for Knowledge Catalog

Database *

Host name or IP Address *

Port *

Port is SSL-enabled The port is configured to accept SSL connections

SSL Certificate

Connection discovery Discover data assets

Project for discovered assets *

Credentials
 Personal Shared

All project collaborators use the provided credentials for the connection.

User name *

Password *

Enter information for the selected data source

Screenshot

Most of connection information is pre-populated based on the global connection

- Click the **Discover data assets** check box under Connection discovery.
- Select the **Auto Discovery<id>** project from the “Project for discovered assets” selection list.
- Choose **Shared** for Credentials
- Click **Test**
- Click the **Create** button.

The screenshot shows the 'Auto Discovery' project page in IBM Cloud Pak for Data. The 'Assets' tab is highlighted with a red arrow. In the top right, there are counts for 'Assets' (3) and 'Collaborators' (1). The 'Recent activity' section displays two log entries:

- Discovery process has completed for connection Db2 Travel to project Auto Discovery 3:54 PM
- Discovery process has started for connection Db2 Travel to project Auto Discovery 3:54 PM

The auto discovery process discovered 3 data assets and added them to the project. Your discovery may be different from this screenshot because this connection is a shared Db2 database, and data assets are being added and removed all the time by other users. Knowledge Catalog scanned the Db2 database instance and collected the metadata for all the user data assets that user is authorized to access. The **Recent activity** area notifies the status of recent discovery processes.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED
<input type="checkbox"/>	RENTALCAR	Data Asset	admin	6 Apr 2020, 11:49:18 pm
<input type="checkbox"/>	HOTEL	Data Asset	admin	6 Apr 2020, 3:54:30 pm
<input type="checkbox"/>	Db2 Travel	Connection	icp4d-dev	6 Apr 2020, 3:54:28 pm

- Click the **Assets** tab to view the discovered assets.
- Click on the **HOTEL** data asset to open the data previewer.

Schema: 8 Columns
Preview: 394 rows Last refresh: 4 minutes ago

BOROUGH	PRICE	NAME	SERVICE	NUM_RATING	RATING	LOCATION	ACCOUNT_NUM
Type: String	Type: Integer	Type: String	Type: String	Type: Integer	Type: Integer	Type: String	Type: Smallint
BOROUGH		NAME	SERVICE			LOCATION	0
Queens	64	Flushing Ymca	Free WI-Fi;Swimming F	69	3	Queens	0
Manhattan	72	Harlem YMCA	Free WI-Fi;24-hour Fro	223	3	Manhattan	0
Brooklyn	74	NEW WORLD HOTEL	Free WI-Fi;24-hour Fro	414	1	Brooklyn	0
Manhattan	86	Vanderbilt YMCA	Free WI-Fi;Swimming F	1525	3	Manhattan	0
Queens	95	Lexington Inn Queens	Free WI-Fi;Business Ce	260	3	Queens	0
Brooklyn	98	Canal Loft Hotel	Concierge Service;Lugg			Brooklyn	0
Queens	102	Kamway Lodge	Free WI-Fi;Family Roor	9	3	Queens	0
Queens	104	Flushing Hotel		69	2	Queens	0
Brooklyn	107	Hostel - Chrystie Street	Free WI-Fi;Internet Se			Brooklyn	0
Manhattan	108	Hotel 309	24-hour Front Desk;Ga	492	3	Manhattan	0
Queens	108	The Parc Hotel	Free WI-Fi;24-hour Fro	281	4	Queens	0
Brooklyn	110	Grandview Hotel Brook	Free Internet Services;	17	3	Brooklyn	0
Queens	116	Flushing Central Hotel	Free WI-Fi;24-hour Fro	168	3	Queens	0

The information panel on the right shows a tag of **DB2INST1**. This is the schema in the Db2 instance that the table came from. The top left section provides information about the total number of columns and rows in the table. In the data section, scroll to the right, to see all columns.

- Click on the **Profile** tab to get frequency and summary statistics for each of your columns. It's a quick way to get an understanding of the data by look at the metrics of data distribution.
- If you are running profile first time on the table, click on **Create Profile**. It will take few minute depends on the volume of data in the table.
- The **Refine** option on left (under **Preview** tab), let you use the Data Refinery tool against the data assets.

Current profile		Last profile		Columns	Rows	Delete	Update Profile
166 classifiers		7 Apr 2020 - 12:29 am View Log		8	394		
BOROUGH Type: Varchar		PRICE Type: Integer		NAME Type: Varchar	SERVICE Type: Varchar		
• City 99% 1% 0%		• Income 98% 2% 0%		• Person Name 0% 100% 0%	• Text 0% 92% 8%		
FREQUENCY Manhattan Brooklyn Queens Staten Island Central Bayside Long Beach Suburban BOROUGH		FREQUENCY 193 - 194 230 - 231 228 - 229 220 - 221 218 - 219 212 - 213 210 - 211 226 - 227 185 - 186		FREQUENCY Arlo SoHo The Bryant Park Hotel Brooklyn Apartments Radisson Hotel Queens Ramada by Wyndham Queens City The Kimpton Muse Hotel Holiday Inn Manhattan-Financial District The Lex NYC Club Quarters Hotel Wall Street Marco LaGuardia Hotel	FREQUENCY Free WI-Fi;24-hour Front Desk;Multilingual Sta... Free Internet Services;Free WI-Fi;24-hour Fro... Missing 24-hour Front Desk;Multilingual Staff;Restaura... Free WI-Fi;24-hour Front Desk;Restaurant; Free WI-Fi;24-hour Front Desk;Terrace; Free WI-Fi;Family Rooms;Internet Services; Free WI-Fi;24-hour Front Desk;Express Check-... 24-hour Front Desk;Multilingual Staff;Concierg... Free WI-Fi;24-hour Front Desk;Parking;		
Showing 9 of 9		Showing 10 of 201 View All		Showing 10 of 100 View All	Showing 10 of 67 View All		
STATISTICS		STATISTICS		STATISTICS	STATISTICS		
Unique		136		Unique	392	Unique	37
Minimum Length		64		Minimum Length	4	Minimum Length	7
Maximum Length		13		Maximum	60	Maximum Length	67

The profile gives an insight of data. It will direct you through the data cleaning requirement by provide frequency and statistics details of each column. Profile can help you to determine if the data refinery flow going in the right direction. For example: BOROUGH column shows most of hotels listed in Manhattan. The NAME column gives an idea of number of unique hotels are there.

8.5 Add Structured Data to Catalog

You have cataloged unstructured data files from the local file system. Now auto discovered data assets from a Db2 database connection. You will now catalog two tables from the Db2 database connection; **Rentalcar** and **Hotel** using the **Connected asset** catalog method. These tables are needed for build an application on CPD.

Go to the **Navigation menu** in the top left corner and click on **Travel<id>** catalog from **Organize > All catalogs**.

Your catalogs

Admin	Travel	→ ⋮ → View Delete
Creator:	admin	Travel Knowledge Catalog
Date created:	Apr 04, 2020 2:11 PM	

- Click on the **configuration** icon of the **Travel<id>** catalog.
- Select **View**

Catalogs > Travel

Add to Catalog ▾

- Local files
- Connected asset
- Connection

- Click **Add to Catalog > Connected asset** from the Catalog menu.

Source* Select Source

Name* Travel Hotel

Description Hotel details from New York City

Business Terms Search Business Terms

Tags Start typing to add values +
Travel X

Classification* None

Privacy ● Public ○ Private
All catalog members can find and use the asset.

Members + Add members

Cancel Add

- Enter a Name of **Travel Hotel<id>**.
- Enter a Description of **Hotel details from New York City**.

- Click in the **Tags** area and select the **Travel** tag from the drop-down list.
- Click the **Select Source** button to choose a Connection to add connected assets from.

The screenshot shows the IBM Cloud Pak for Data interface. In the top navigation bar, 'Catalogs' is selected. Below it, the path 'Travel' is shown. The main content area displays a tree view of database objects:

- Connections:** Db2 Travel (highlighted)
- Schemas:** DB2INST1 (highlighted)
- Tables:** HOTEL (highlighted), RENTALCAR

At the bottom right of the interface, there are two buttons: 'Cancel' and 'Select' (which is highlighted with a red border).

- Click on the **Db2 Travel** connection from the list of connections.
- Click on the **DB2INST1** schema from the list of schemas.
- Click on the **HOTEL** table from the list of tables.
- Click the **Select** button.

On the following window click the **Add** button.

You should see the catalogued table as a data asset with the tag you supplied.

Similar way adds next table to catalog.

The screenshot shows the 'Add to Catalog' menu options:

- Local files
- Connected asset (highlighted with a red border)
- Connection

- Click **Add to Catalog > Connected asset** from the Catalog menu.

The screenshot shows the 'Add asset from connection' form:

- Source***: Select Source (highlighted with a red border)
- Name***: Travel Rentalcar
- Description**: Rental car details from New York City
- Business Terms**: Search Business Terms
- Tags**: Start typing to add values (Travel X)
- Classification***: None
- Privacy**: Public (radio button selected)
- Members**: Add members

- Enter a Name of **Travel Rentalcar<id>**.
- Enter a Description of **Rental car details from New York City**.
- Click in the **Tags** area and select the **Travel** tag from the drop-down list.
- Click the **Select Source** button to choose a Connection to add connected assets from.

The screenshot shows the IBM Cloud Pak for Data Catalogs interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. Below it, the breadcrumb path shows 'Catalogs > Travel'. The main area displays a hierarchical tree of database connections:

Connections	Db2 Travel	DB2INST1
Connections (1)	Schemas (1)	Tables (2)
Db2 Travel	> DB2INST1	> HOTEL
		RENTALCAR

At the bottom right of the interface, there are 'Cancel' and 'Select' buttons, with 'Select' being highlighted with a red box.

- Click on the **Db2 Travel** connection from the list of connections.
- Click on the **DB2INST1** schema from the list of schemas.
- Click on the **RENTALCAR** table from the list of tables.
- Click the **Select** button.

On the following window click the **Add** button.

You should see the catalogued table as a data asset with the tag you supplied.

9 Discover Assets using Quick Scan

A quick scan can come to great help when you don't know your data very well, and you want to analyze large data sets to see a general overview of the quality of the data. Quick scan performs analyze columns, analyze data quality and automatically assign term.

In this task you will use the **Db2 Travel** global connection to **Travel** database on the Db2.

Go to the **Navigation menu** in the top left corner and click on **Organize > Curation > Data discovery**. Select **Auto Discovery** from list of projects.

The screenshot shows the 'Data discovery' section of the IBM Cloud Pak for Data interface. At the top, there are navigation links for 'IBM Cloud Pak for Data', 'All', and 'Search'. Below that, a 'Data discovery' section header is visible. Underneath, a 'Quick scan results' card is shown with the following details:

- Summary** tab is selected.
- Pending analysis**, **Action required**, and **Reviewed** tabs are also present.
- 8 total jobs** are listed.
- 63%** completion progress bar.
- New discovery job** button (highlighted with a red arrow).
- View workspaces** link.
- Quick scan** and **Automated discovery** options with their descriptions and 'Learn more' links.

- Click **New discover job** > **Quick scan**

The screenshot shows the 'Quick scan job' configuration page. At the top, there are navigation links for 'IBM Cloud Pak for Data', 'All', and 'Search'. Below that, a 'Quick scan job' section header is visible. Underneath, a 'Connection' dropdown is set to 'Add a connection' (highlighted with a red box). A red arrow points to the 'Add a connection' button. Other connection options listed include 'mortgage-db2', 'DV', and 'hive-hdp3.1-wkc2'. A 'Browse' button is also visible.

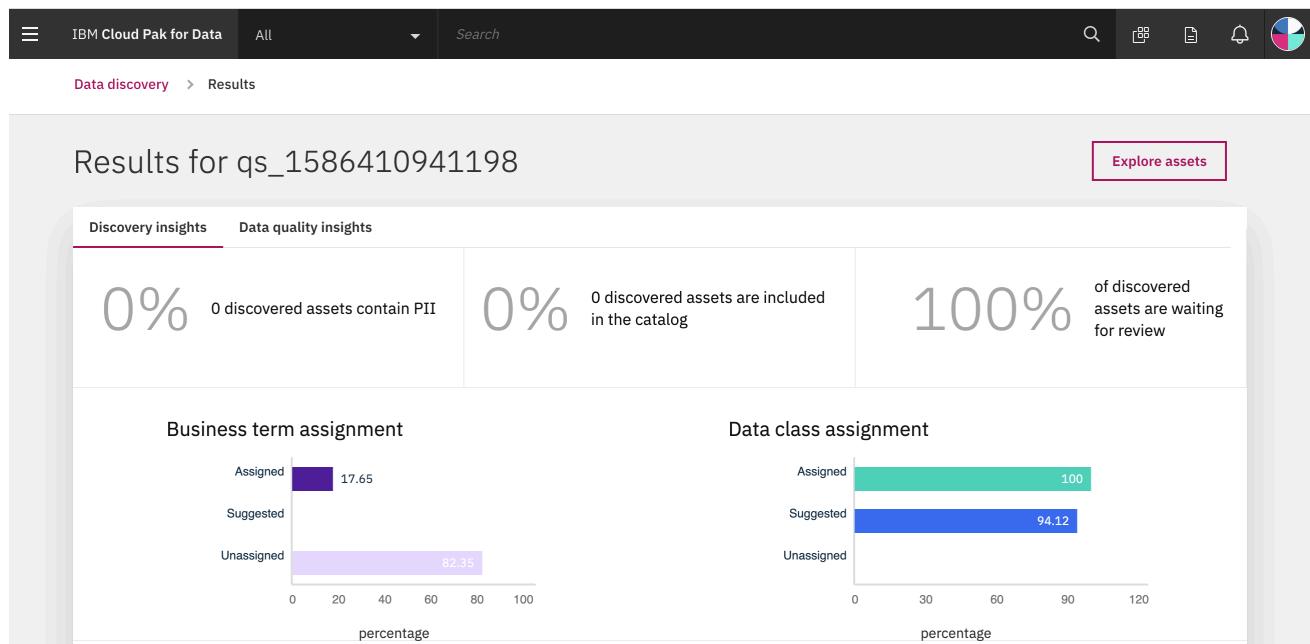
- Click on **Add a connection**.
- Choose connection name **Db2 Travel**.
- Chick **Next** button.

The screenshot shows the 'Quick scan job' configuration page. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. The main form includes fields for 'Connection *' (set to 'Db2 Travel'), 'Discovery root' (set to 'schema[TRAVEL|DB2INST1]'), 'Discovery options' (checkboxes for 'Analyze columns', 'Analyze data quality', 'Assign terms', and 'Use machine learning to assign terms' are checked), and a 'Browse' button. Below these are settings for 'The maximum number of records included in the data set sample:' (set to 1000). A 'Workspace *' section shows 'Travel Workspace' selected. At the bottom right are 'Cancel' and 'Discover' buttons.

- Click on **Add a connection**
- Choose the connection named **Db2 Travel** that you created previously, click Next
- Select Discover root as **TRAVEL > DB2INST1**
- Check necessary Discover options
- Click on Add a workspace under Workspace and named it as **Travel Workspace**. Click Create.
- Click on **Discover**

The screenshot shows the 'Quick scan results' page. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. The main area displays 'Data discovery' results. A table lists quick scan jobs, with one row selected (Job ID: qs_1586410941198, Data assets: 2, Connection: Db2 Travel, Started by: admin, Processing time: 30 seconds, Status: Ready for review, Status updated: April 9, 2020 12:42 AM). The table has tabs for 'Summary', 'Pending analysis', 'Action required' (which is active), and 'Reviewed'. There are also buttons for 'View results', 'Discover again', and 'Delete'. A status filter on the left shows 'All jobs requiring action' is selected. At the top right, there are links for 'New discovery job', 'View workspaces', and 'View automated discovery results'.

- Click on **Quick scan results**
- Choose **Action required** tab
- Check the quick scan result associated with **Db2 Travel** connection
- Click **View results** tab to explore the discover assets.



- Review the discovery results using **Explore assets** tab

Review assets for proper business term assignment, if needed you can adjust them.

The screenshot shows the 'Assets' page for asset 'qs_1586410941198'. On the left, there are filters for Asset type (File, Schema, Table, Column), Labels (No filters of this type), Tables (RENTALCAR, HOTEL selected), and Schemas. Buttons for 'Clear' and 'Apply' are present. The main area displays a table of 'Discovered columns (17)'. The columns include Column name, Identity, Quality, Assigned business term, Suggested business term, Assigned data class, Suggested data class, and term actions. Examples shown include ACCOUNT_NUM (Boolean 100%, Account Number 100%), HOTEL (Boolean 100%, Account Number 100%), and RENTALCAR (Code 100%, Middle Name 50%). A search bar and pagination controls (1 of 2 pages, 1-10 of 17 items) are at the bottom.

- Select Asset type as **Column**
- Filters necessary tables **RENTALCAR, HOTEL** using checkbox
- Click on **Apply**

The screenshot shows the IBM Cloud Pak for Data interface. The top navigation bar includes 'IBM Cloud Pak for Data', 'All', 'Search', and various icons. Below the navigation is a breadcrumb path: 'Data discovery > Results > Assets'. The main area is titled 'Assets for qs_1586410941198'. On the left, there's a sidebar for 'Asset type' (File, Schema, Table, Column) and 'Filters'. The main content area has three buttons: 'Approve results', 'Reject results', and 'Audit assets'. It displays two selected items: 'HOTEL' and 'RENTALCAR'. Both entries show 'Ready for review' status. The table columns include: Table name, Identity, Quality, Schema name, Discovery root, Status, Connection name, and Workspace.

	Table name	Identity	Quality	Schema name	Discovery root	Status	Connection name	Workspace
<input checked="" type="checkbox"/>	HOTEL	HOTEL- qs_1586410941198	96%	DB2INST1	schema[TRAVEL DB2IN	Ready for review	Db2 Travel	Travel Workspace
<input checked="" type="checkbox"/>	RENTALCAR	RENTALCAR- qs_1586410941198	98%	DB2INST1	schema[TRAVEL DB2IN	Ready for review	Db2 Travel	Travel Workspace

- Change Asset type as **Table**.
- Select all **Travel** related tables.
- Click on **Approve results**.
- Click on **Approve assets**.

10 Implement Business Glossary

Cloud Pak for Data enables you to structure your enterprise information in a logical way, discover relationships between assets, and keep your data always up-to-date. You can import existing glossary with categories, terms, information governance policies and rules.

In this task, you will import some pre-generated business glossaries, apply rules and polices to govern data.

10.1 Download Business Glossaries

First download pre-generated business glossaries from the GIT to your local machine.

- Go to <https://github.com/sanjitc/TravelBid/tree/master/assets/BusinessGlossary>

File	Description	Created
travel-business-glossary-term.csv	Create travel-business-glossary-term.csv	5 minutes ago
travel-business-glossary-category.csv	Create travel-business-glossary-category.csv	5 minutes ago
travel-business-glossary-term.csv	Creating travel-business-glossary-term.csv file	4 minutes ago

- Download two CSV files and save locally.

Notes:

- Go to that particular business glossaries that you want to **download** and click on it.
- You will see "Raw" button on the top right side of the dataset.
 - Press "Alt" and then left click the "Raw" button (on Windows) or
 - Click with two fingers (on Mac)
- Download link file

10.2 Import categories

Categories provide the logical structure for governance artifacts so that you can browse and understand the relationships among artifacts. Categories can be organized in a hierarchy based on their meaning and relationships to one another. You can modify governance artifacts outside of the catalog and import them from a file that is in CSV format. Sequence is important when importing business glossaries. Make sure import categories before do the terms.

Go to the **Navigation menu** in the top left corner and click on **Organize > Data and AI governance > Categories**.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. Below the navigation bar, a progress bar indicates the steps: 'Choose file' (done), 'Set merging' (not done), and 'Import' (not done). A red arrow points to the 'Import' button.

- Click **Import** to import the CSV file contains category information that you downloaded from Git.



Choose file

Must be a CSV file.

Add file

The CSV file must conform to the template for importing governance artifacts.
[Learn more](#)

- Click **Add file** and browse the CSV that contains category information.



Choose file

Must be a CSV file.

travel-business-glossary-category.csv

The CSV file must conform to the template for importing governance artifacts.
[Learn more](#)

- Choose the **travel-business-glossary-category.csv** file location
- Click **Next**



Select merge option

Replace all values

Imported values in the CSV file replace existing values in the catalog.

Replace with defined values

Imported CSV values that are not empty replace existing values in the catalog.

Replace empty values

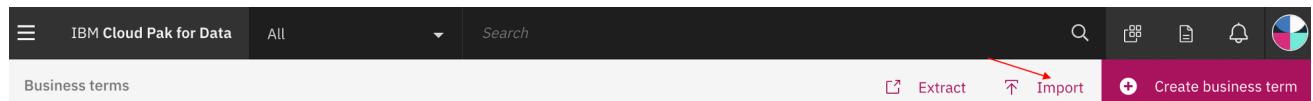
Imported values in the CSV file replace only empty values in the catalog.

- Select merge option as **Replace all values**.
- Click **Import**.

10.3 Import Terms

You create business terms to standardize definitions of business concepts so that your data is described in a uniform and easily understood. You can assign one or more business terms to individual columns in relational data sets, to other governance artifacts, or to data assets.

Go to the **Navigation menu** in the top left corner and click on **Organize > Data and AI governance > Business terms**.



- Click **Import** to import the CSV file contains terms information that you downloaded from Git.
- Next click **Add file** and browse the CSV that contains term information.



Choose file

Must be a CSV file.

[travel-business-glossary-term.csv](#)

The CSV file must conform to the template for importing governance artifacts.

[Learn more](#)

- Choose the **travel-business-glossary-term.csv** file location
- Click **Next**



Select merge option

Replace all values

Imported values in the CSV file replace existing values in the catalog.

Replace with defined values

Imported CSV values that are not empty replace existing values in the catalog.

Replace empty values

Imported values in the CSV file replace only empty values in the catalog.

- Select merge option as **Replace all values**.
- Click **Import**.
- On import term completion, click **Go to task**.

Assigned tasks Completed tasks

Sort by: Due date ▾

Task name	Workflow status	Due date	Task created
Publish Business terms ⌚ No due date 3 minutes ago	Not started	⌚ No due date	3 minutes ago

1 item selected **Cancel**

- Select the check box for **Publish Business terms** that you just imported
- Click **Publish** tab

10.4 Automated Discovery

Discover assets to add data to the default catalog. During the discovery the data is imported, analyzed, and classified.

Earlier you ran quick scan, but in this task you will re-run discover assets, so data can imported, analyzed, and classified according the glossary you imported/created earlier.

Go to the **Navigation menu** in the top left corner and click on **Organize > Curation > Data discovery**.

New discovery job ▾ ⌚ View workspaces ⌚ View automated discovery results

Quick scan
Run analysis of data sample and skip data import.
[Learn more](#)

Automated discovery.
Run an in-depth analysis of all assets and import the assets.
[Learn more](#)

9 total jobs 44% of job

- Navigate to **New discover job > Automated discovery**

The screenshot shows the 'Automated discovery job' configuration page. At the top, there is a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. The main section is titled 'Automated discovery job'. It includes fields for 'Connection *' (set to 'Db2 Travel'), 'Discovery root' (set to 'schema[TRAVEL|DB2INST1]'), and 'Discovery options' (checkboxes for Analyze columns, Analyze data quality, Assign terms, Use machine learning to assign terms, and Use data sampling, all checked). A text input field for 'The maximum number of records included in the data set sample:' contains the value '1000'. Below this, a 'Workspace *' dropdown is set to 'Travel Workspace'. At the bottom right are 'Cancel' and 'Discover' buttons.

- Choose the connection named **Db2 Travel** that you created previously, click Next
- Select Discover root as **TRAVEL > DB2INST1**
- Check necessary Discover options
- Click on Add a workspace under Workspace and named it as **Travel Workspace**. Click Create.
- Click on **Discover**

The screenshot shows the 'Discovered assets' page. It displays a table with the following data:

Asset name	Asset type	Tables	Status	Actions	
DB2INST1	Schema	2	Phase: Import Finished Start: April 10, 2020, 1:43 AM End: April 10, 2020, 1:43 AM	Phase: Analyze Running Start: April 10, 2020, 1:43 AM Done: 0% Successful 0% Cancelled 0% Failed 0%	Review

At the top right of the table, there are 'Refresh' and 'Actions' buttons. Red arrows point from the text 'Refresh icon' to the refresh button and from 'Review icon' to the review button.

- Discovery job will take some time to complete. Use the **refresh icon** to check the status of the job.
- User **review icon** to discovers how terms automatically assigned to data sets.

11 Data Assets Sentiments

As data assets are cataloged, they are automatically profiled and classified so data consumers can have a better understanding of their content. They can then be enriched using Knowledge Catalog's social capabilities.

In this task, you will visit the **Profile** section of a structured and unstructured data asset, to examine the different profiling and classification features provided. You will also visit the **Review** section to experience how you can rate and review assets to allow others to easily identify and evaluate them based on their ranking and comments.

Go to the **Navigation menu** in the top left corner and click on **Travel<id>** catalog from **Organize > All catalogs**.

Your catalogs

Admin

Travel

Creator: admin Date created: Apr 04, 2020 2:11 PM

Travel Knowledge Catalog

New Catalog

... View Delete

- Click on the **configuration** icon of the **Travel<id>** catalog.
- Select **View**

IBM Cloud Pak for Data

All

Search

Catalogs > Travel

What assets are you looking for?

Any type | Any source | Any tag | Clear all

Watson Recommends | Highly Rated | Recently Added | Collapse

Watson Recommends	Highly Rated	Recently Added	
Data asset Travel Rentalcar	Data asset Travel Hotel	Connection Db2 Travel	Data asset Knowledge-Ggra
Owner: admin Added: Apr 07, 2020 11:45 PM..PM Tags: Travel	Owner: admin Added: Apr 07, 2020 11:29 PM..PM Tags: Travel	Owner: admin Added: Apr 06, 2020 3:54 PM..PM Tags: Db2 Travel	Owner: admin Added: Apr 04, 2020 7:2 Tags: Docu... Travel
☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews	☆☆☆☆☆ 0 reviews

- Click on the **Recently Added** section of the data asset browser.
- Click on the **Travel Rentalcar<id>** asset to view its properties.

IBM Cloud Pak for Data All Search

Catalogs > Travel > Travel Rentalcar

Add to Catalog

Overview Access Review Profile Lineage

DATA ASSET

Travel Rentalcar

Remove Download Add to Project

Description		Schema: 9 Columns 297 Rows								
Rental car details from New York City		Preview: 297 rows Last refresh: 1 minute ago Refresh								
Added:	Apr 07, 2020 11:45 PM..PM	MODEL	SIZE	RENT	AGENCY	AGENCYPH...	ACCOUNT_...	SIZEID		
Format:	application/octet-stream	Type: String	Type: String	Type: Integer	Type: String	Type: String	Type: Smallint	Type: Smallint		
Size:	77 KB	Text	Code	Code	Text	US Phone N...	Boolean	Code		
Business Terms		Chevrolet Spark	Economy	73	Alamo	332 W 44th St. N	8888266893	0	1	
There are no terms available for this asset.		Nissan Versa	Compact	86	Alamo	332 W 44th St. N	8888266893	0	2	
Tags		Toyota Corolla	Midsize	69	Alamo	332 W 44th St. N	8888266893	0	3	
Travel		Ford Fusion	Fullsize	76	Alamo	332 W 44th St. N	8888266893	0	8	
Reviews		Toyota Avalon	Premium	82	Alamo	332 W 44th St. N	8888266893	0	9	
Db2 Travel		Cadillac ATZ	Luxury	85	Alamo	332 W 44th St. N	8888266893	0	10	
Db2		Dodge Grand Car	Minivan	92	Alamo	332 W 44th St. N	8888266893	0	6	
type:		Ford Mustang	Convertible	282	Alamo	332 W 44th St. N	8888266893	0	11	

The **Overview** section of the asset where you can view sample rows and metadata about the asset, including column level classifications, if it's a data asset. You can modify its name and description, add tags and assign business terms and classifications at the asset or column level.

- Click on the **Profile** section of the data asset.

The profile should automatically appear. If not, and you are presented with a method to create or update the profile, follow the instructions to do so.

IBM Cloud Pak for Data All Search

Catalogs > Travel > Travel Rentalcar

Add to Catalog

Overview Access Review **Profile** Lineage

Current profile		Last profile		Columns	Rows	Delete	Update Profile
166 classifiers		7 Apr 2020 - 11:45 pm View Log		9	297		
BOROUGH Type: Varchar		MODEL Type: Varchar		SIZE Type: Varchar	RENT Type: Integer		
<input checked="" type="radio"/> City 89% 11% 0%		<input checked="" type="radio"/> Text 0% 100% 0%		<input checked="" type="radio"/> Code 0% 100% 0%	<input checked="" type="radio"/> Code 0% 100% 0%		
FREQUENCY 		FREQUENCY 		FREQUENCY 	FREQUENCY 		
<i>Showing 5 of 5 </i>							

The profile of a data asset that contains relational or structured data, shows information about each column in the data set, based on the first 5,000 rows of data. The profile shows the frequency of the inferred attribute classifiers and statistics about the data for each column.

- Scroll down and to the right to review all the profiling statistics.
- Click the **Review** section to rate and review the data asset.

- Copy and Paste the following bolded text into the Description: **This travel rentalcar data is quality data that comes from different rental car agencies. However, in order to get the full value of this data it needs to be combined with the auto insurance hotel data.**
- Click the **4th star** from the left to give the asset a 4 stars rating.
- Click the **Submit** button.

Now you should have one review with an Overall Rating of 4.0.

Let's take a look into one of the documents in the **Travel** catalog. Click on the **Building_Travel_Reservation_Application_with_CPD** data asset that tag as a document to view the properties. As this is a PDF asset, you should able to see the document also. **Scroll** down to view the content of the document.

Click the **Review** section to rate and review the data asset.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below it, a breadcrumb trail shows 'Catalogs > Travel > Building_Travel_Reservation_App...'. On the right, there are icons for 'Add to Catalog' and notifications.

The main content area is titled 'Building_Travel_Reservation_Application_with_CPD'. It has tabs for 'Overview', 'Access', 'Review' (which is selected), and 'Lineage'. On the left, there's a sidebar with 'Overall Rating' (0.0, 5 stars) and a 'Review Summary' section showing counts for 0, 1, 2, 3, 4, and 5 reviews. The main panel is titled 'My Review' and shows a review by 'admin' from April 08, 2020, with a 3-star rating. The review text is: 'An interesting document that shows how to achieve the most basic data tasks in combination of CPD and Shiny R package: in the context of collect, organize, virtualize and infuse data assets to build a web page.' At the bottom right of this panel are 'Cancel' and 'Submit' buttons.

- Copy and Paste the following bolded text into the Description: **An interesting document that shows how to achieve the most basic data tasks in combination of CPD and Shiny R package: in the context of collect, organize, virtualize and infuse data assets to build a web page.**
- Click the **3rd star** from the left to give the asset a 3 stars rating.
- Click the **Submit** button.

12 Shop for Data

In this task, you will leverage Knowledge Catalog's intelligent **Shop for Data** AI-powered **Search and Suggest** experience that guides you to the most relevant assets in the catalog, based on understanding of relationships between assets, usage of those assets and social connections between the users of those assets.

You will also use the **Filter** section of the Knowledge Catalog that is automatically built and **Organized** by *Asset Type* and *Tag* as you catalog assets. Tagging is essential when cataloging assets, it expedites the process for consumers to easily search and find what they are looking for.

Name	Owner	Tags	Business Terms	Type	Date Added
Building_Travel_Reservation_Application_with...	admin	Travel Document		Data asset	Apr 04, 2020
Db2 Travel	admin	Db2 Travel		Connection	Apr 06, 2020
Knowledge-Ggraph-With-WKC	admin	Travel Document		Data asset	Apr 04, 2020
Travel Hotel	admin	Travel		Data asset	Apr 07, 2020
Travel Rentalcar	admin	Travel		Data asset	Apr 07, 2020

12.1 Shop Data Assets using Suggestion

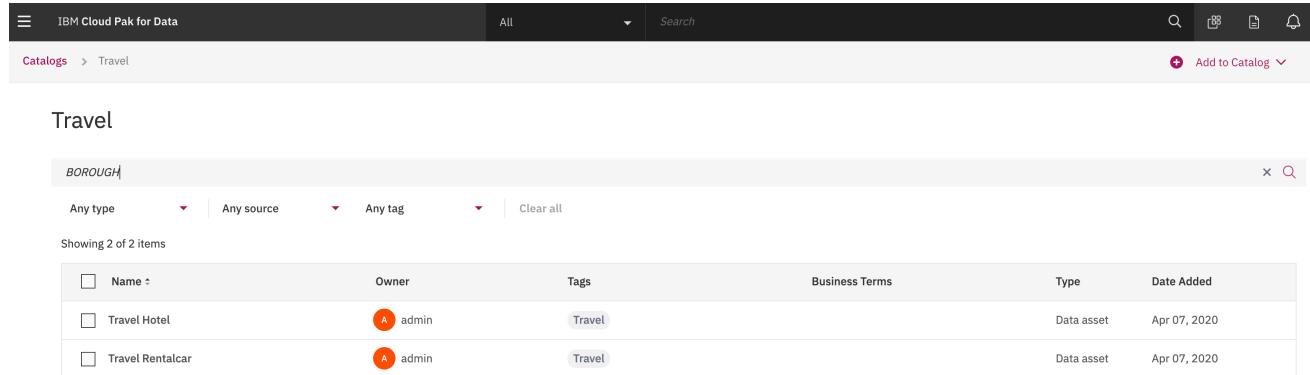
You can easily search for data using the **Watson Recommends**, **Highly Rated** and **Recently Added** suggestion categories to find relevant data. These categories are automatically populated by Knowledge Catalog as you catalog, curate and enrich data assets. For example: Highly Rated assets displays previously reviewed assets.

Name	Owner	Tags	Business Terms	Type	Date Added
Travel Rentalcar	admin	Travel		Data asset	Apr 07, 2020 11:45 PM..PM
Building_Travel_Reservati...	admin	Docu... Travel		Data asset	Apr 04, 2020 7:29 PM..PM
Knowledge-Ggraph-With...	admin	Travel Docu...		Data asset	Apr 04, 2020 7:29 PM..PM
Db2 Travel	admin	Db2 Travel		Connection	Apr 06, 2020 3:54 PM..PM
Travel Hotel	admin	Travel		Data asset	Apr 07, 2020 11:29 PM

- Click the **Travel** link at the top of the page to go back to the catalog asset browser.
- Click the **Highly Rated** section.

12.2 Shop Data Assets using Search

In this section you will shop for data by specifying search criteria using the **Search area** of the Knowledge Catalog asset browser. The search criteria are not case sensitive. You search criteria could be a tag, column name, asset name, business terms, description etc.



The screenshot shows the IBM Cloud Pak for Data Knowledge Catalog interface. The search bar at the top contains the text "BOROUGH". Below the search bar, there are filters for "Any type", "Any source", and "Any tag", each with dropdown menus. A "Clear all" button is also present. The results table shows two items found, both labeled "Travel".

Name	Owner	Tags	Business Terms	Type	Date Added
Travel Hotel	admin	Travel		Data asset	Apr 07, 2020
Travel Rentalcar	admin	Travel		Data asset	Apr 07, 2020

- Inside the Knowledge Catalog search area, type in **BOROUGH**.
- Data assets are displayed that have a tag, column name, asset name or description that contains the consecutive letters of **borough**.
- Click the **x** at the far right of the search area to clear the search.

13 Prepare Data for Analytics Project

Let's talk about the use case to have a better understanding on the analytics project. Imagine TravelBid is a startup, offering discounted hotel and rental car reservations in New York City. They have a bulk of hotel and rental car vendors, from where they fulfill travel reservation for their clients. TravelBid needs to quickly build an e-commerce platform where clients can "name their own price". Depends on client's named price, system will generate list of hotel and car matches within the prices. Plan here to build an interactive GUI based application to address TravelBid's requirement using Shiny R Package on CPD, accessing tables from different relational tables. You noticed earlier there are two travel related data sources. 1) The **HOTEL** table that stores information about hotels in New York City. 2) The **RENTALCAR** table that stores information about different rental car agencies and fleets. Schema of these two tables are as follows:

Table 1: HOTEL			Table 2: RENTALCAR		
<input type="checkbox"/>	Column Name	Data Type	<input type="checkbox"/>	Column Name	Data Type
<input checked="" type="checkbox"/>	ACCOUNT_NUM	SMALLINT	<input checked="" type="checkbox"/>	ACCOUNT_NUM	SMALLINT
<input checked="" type="checkbox"/>	BOROUGH	VARCHAR	<input checked="" type="checkbox"/>	AGENCY	VARCHAR
<input checked="" type="checkbox"/>	LOCATION	VARCHAR	<input checked="" type="checkbox"/>	AGENCYADDR	VARCHAR
<input checked="" type="checkbox"/>	NAME	VARCHAR	<input checked="" type="checkbox"/>	AGENCYPHONE	VARCHAR
<input checked="" type="checkbox"/>	NUM_RATIN	INTEGER	<input checked="" type="checkbox"/>	BOROUGH	VARCHAR
<input checked="" type="checkbox"/>	PRICE	INTEGER	<input checked="" type="checkbox"/>	MODEL	VARCHAR
<input checked="" type="checkbox"/>	RATING	INTEGER	<input checked="" type="checkbox"/>	RENT	INTEGER
<input checked="" type="checkbox"/>	SERVICE	VARCHAR	<input checked="" type="checkbox"/>	SIZE	VARCHAR
			<input checked="" type="checkbox"/>	SIZEID	SMALLINT

In this task, you will use the **IBM Data Refinery** to shape data that you cataloged in the **Travel** catalog. You will gain an understanding of the data preparation capabilities of the Data Refinery and how it can help you understand and visualize data in preparation for analytics project.

13.1 Add Catalogued Data Assets to Analytics Project

In order to refine data, the data needs to be in a project. You will add **HOTEL** and **RENTALCAR** data assets from the **Travel** catalog to the Travel project to prepare it for analytics and AI.

Go to the **Navigation menu** in the top left corner and click on **Travel** catalog from **Organize > All catalogs**.

Your catalogs

Travel

Creator: admin Date created: Apr 04, 2020 2:11 PM

Travel Knowledge Catalog

New Catalog

⋮

View

Delete

- Click on the **configuration** icon of the **Travel** catalog.
- Select **View**

IBM Cloud Pak for Data

All

Search

Catalogs > Travel

Add to Catalog

Browse Assets Access Control Settings

Travel

What assets are you looking for?

Type: Data asset, Source: Db2 Travel

Add to Project Remove

2 of 2 items selected Cancel

Name	Owner	Tags	Business Terms	Type	Date Added
Travel Hotel	admin	Tr...	Sensitive Information	Data asset	Apr 10, 2020
Travel Rentalcar	admin	Tr...		Data asset	Apr 10, 2020

- Click the **check box** next to the **Travel Hotel** data asset.
- Click the **check box** next to the **Travel Rentalcar** data asset.
- Click the **check box** next to the **DB2 Travel** connection.
- Click the **Add to Project** link at the top of the data asset list.

Add to Project

Target*

Travel

Selected assets (2)

Asset Name	Catalog	Connection
Travel Hotel	Travel	Db2 Travel
Travel Rentalcar	Travel	Db2 Travel

Connections to be added (1)

Db2 Travel

Travel
Connect to Db2 Travel database for Knowledge Catalog

Add

- Select the **Travel** project from the list of Target projects.
- Click the **Add** button.

13.2 Refine Data

Go to the **Navigation menu** in the top left corner and chose **Projects > Travel**

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	Travel Rentalcar	Data Asset	admin	13 Apr 2020, 10:35:15 am	
<input type="checkbox"/>	Travel Hotel	Data Asset	admin	13 Apr 2020, 10:35:15 am	
<input type="checkbox"/>	Db2 Travel	Connection	admin	13 Apr 2020, 10:35:14 am	

- First test the **Db2 Travel** connection. You should about connect to the database, if not check the username and password.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	Travel Rentalcar	Data Asset	admin	13 Apr 2020, 10:35:15 am	
<input type="checkbox"/>	Travel Hotel	Data Asset	admin	13 Apr 2020, 10:35:15 am	
<input type="checkbox"/>	Db2 Travel	Connection	admin	13 Apr 2020, 10:35:14	Publish to Catalog Refine Download Promote Remove

- Click on the **Assets** tab at the top of the project page.
- Select the ellipses... to the right of the **Travel Rentalcar** data asset to view the data asset action menu.
- Select the **Refine** menu item.

You are brought into the **IBM Data Refinery** to begin shaping the **Travel Rentalcar** data. In the subsequent steps, you will use some of the Data Refinery operations to shape the rentalcar data you added to the project and create a newly shaped dataset that you will put back to the project as a **CSV** file that will be used for analytics and AI.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. Below the navigation is a breadcrumb trail: 'My Projects > Travel > Travel Rentalcar > Refine data'. On the left, there's a table preview with columns: BOROUGH, MODEL, SIZE, and RENT. The table contains 13 rows of data. To the right of the table are sections for 'DATA REFINERY FLOW DETAILS' (Location: Travel, Flow Name: Travel Rentalcar_flow) and 'DATA REFINERY FLOW OUTPUT' (Location: Travel/Data assets, Dataset Name: DATA SET NAME).

	BOROUGH	MODEL	SIZE	RENT
1	Manhattan	Chevrolet Spark	Economy	73
2	Manhattan	Nissan Versa	Compact	86
3	Manhattan	Toyota Corolla	Midsized	69
4	Manhattan	Ford Fusion	Fullsize	76
5	Manhattan	Toyota Avalon	Premium	82
6	Manhattan	Cadillac ATZ	Luxury	85
7	Manhattan	Dodge Grand Caravan	Minivan	92
8	Manhattan	Ford Mustang	Convertible	282
9	Manhattan	Toyota Rav4	Midsized SUV	90
10	Manhattan	Hyundai Santa Fe	Standard SUV	94
11	Manhattan	Chevy Tahoe	Fullsize SUV	99
12	Manhattan	Chevrolet Spark	Economy	73
13	Manhattan	Nissan Versa	Compact	86

SOURCE FILE: Travel Rentalcar SAMPLE SIZE: First 297 rows

- Click in the **Edit** button to edit the Data Flow details

The screenshot shows the 'Edit output' dialog box. It has tabs for 'DATA REFINERY FLOW DETAILS' and 'DATA REFINERY FLOW OUTPUT'. In the 'DATA REFINERY FLOW DETAILS' tab, the location is set to 'Travel'. In the 'DATA REFINERY FLOW OUTPUT' tab, the 'LOCATION' is set to 'Travel/Data assets', the 'DATA SET NAME' is 'Shaped', and the 'DESCRIPTION' is 'Travel data refined and anonymized'. The 'FILE FORMAT' is set to 'CSV'. A checkbox for 'The first line of the file contains column headers' is checked. At the bottom, there's a note: 'Review the Data Refinery flow details and the Data Refinery flow output details.' and a 'Done' button.

- Click the **pencil icon** in the DATA REFINERY FLOW NAME area of the DATA REFINERY FLOW DETAILS section.

- Rename the Data Flow to **Travel Rentalcar Data Flow**.
- Use **Prepare the Travel data for analytics** for the DESCRIPTION field.
- Click the **Apply** button.
- Hover over the **pencil icon** in the LOCATION area of the DATA REFINERY FLOW OUTPUT section.
- Click the **Edit Output** button to change the DATA SET NAME.
- Rename the Data Set to **Travel Shaped<id>**.
- Use **Travel data refined and anonymized** into the DESCRIPTION field.
- Click the **Checked** icon on the toolbar to save the changes.
- Click the **Done** button.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', a search bar, and various icons. Below it, a breadcrumb trail shows 'My Projects > Travel > Travel Rentalcar > Refine data'. The main area has tabs for 'Data', 'Profile', and 'Visualizations'. On the left, a table displays data from a 'Travel Rentalcar' source, with columns for BOROUGH, MODEL, and SIZE. The table has 9 rows of data. To the right, a 'Details' panel is open, showing '0 STEPS' under 'Data Source' and 'Travel' under 'LOCATION'. The 'DATA REFINERY FLOW DETAILS' section shows the name 'Travel Rentalcar Data Flow' and its description 'Prepare the Travel data for analytics'. It also indicates '0 STEPS'.

- Click the **Save** button on the toolbar to save the Data Flow.
- Click the **X** on the Details panel to close the panel and maximize the shaper real estate.

13.3 Combine Data

You will merge Rentalcar and Hotel tables together and create a single data asset for analytics and AI.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data', 'All', and a search bar. Below the navigation is a breadcrumb trail: 'My Projects > Travel > Travel Rentalcar > Refine data'. On the left, a sidebar titled 'Operation' lists various data shaping operations like 'Convert column value to missing', 'Extract date or time value', etc. A red arrow points to the 'Operation' button. In the center, there's a table with columns 'BOROUGH', 'MODEL', and 'SIZE'. To the right, a 'Steps' panel shows '0 STEPS' and a 'Data Source' section with 'Travel Rentalcar'. A red arrow also points to the 'Join' operation in the sidebar.

- Click the **Operation** button to view the shaping operations menu.
- Scroll down and click the **Join** operation.

The screenshot shows the 'Join' operation configuration screen. The top part displays the 'Join' operation details: 'Combine data from two data sets based on a comparison of the values in specified key columns.' It shows an 'Inner join' selected. Below this, a detailed description explains that it returns only matching rows from both datasets. A note states that suffixes will be added to avoid duplicate column names. The 'Source' section lists 'Travel Rentalcar' as the source dataset with a suffix '*X'. The 'Data set to join' section has a 'Add Data Set' button and a field for a suffix '*Y'. The right side shows a preview table with columns 'BOROUGH', 'MODEL', and 'SIZE'. A 'Steps' panel on the right indicates '0 STEPS' and lists the 'Data Source' as 'Travel Rentalcar'.

- Select the **Inner join** method from the join method list.
- Click the **+ Add Data Set** button in the **Data set to join** section.

Data set to join with Travel Rentalcar

Travel	Data assets
Assets (2)	Data assets (2)
Connections >	Travel Hotel
Data assets >	Travel Rentalcar

Data set to join with Travel Rentalcar

Cancel Apply

- Click on the **Data assets** section.
- Select the **Travel Hotel** data asset.
- Click the **Apply** button.

X Operation Code an operation to cleanse and shape your data

Join

The default suffix for each data set will be used to differentiate any duplicate column names in the resulting data set.

Source Data set to join

Travel Rentalcar Travel Hotel

*Suffix *Suffix

_X _Y

JOIN KEYS

Travel Rentalcar	Travel Hotel
BOROUGH	BOROUGH

(+) Add Join Key

Steps

0 STEPS

Data Source

Travel Rentalcar

SOURCE F1: Travel Renta SAMPLE S1: First 297 rows

- Scroll down in the Join properties area until you see the JOIN KEYS section.
- Click in the **Travel Rentalcar** JOIN KEYS column selection list on the left and select the **BOROUGH** column as the join key column.
- Click in the **Travel Hotel** JOIN KEYS column selection list on the right and select the **BOROUGH** column as the join key column.
- Click the **Next** button.

The screenshot shows the IBM Cloud Pak for Data interface. On the left, there's a sidebar with a search bar and a list of frequently used operations: Calculate, Convert column type, Filter, Math, Remove, Rename, Sort ascending, Sort descending, and Substitute. The main area is titled 'Operation' and contains a table with three columns: ONE, ACCOUNT_NUM (Integer), and SIZEID (Integer). The table has 13 rows of data. The right side shows a 'Steps' panel with one step: 'Data Source' (Travel Rentalcar) and 'Join' (JUST ADDED). The 'Join' step is described as 'left-joined data from Travel Hotel based on columns BOROUGH,BOROUGH'.

- Scroll down the column list and keep **checked** the following columns:
- BOROUGH, MODEL, SIZE, RENT, AGENCY, PRICE, NAME, SERVICE, RATING
- Unchecking columns excludes them from the join result.
- Click the **Apply** button.
- Click the **Save** button on the toolbar to save the Data Flow.

13.4 Rename Column

Some of the column names are ironies. Let's change them to a meaningful name.

The screenshot shows the IBM Cloud Pak for Data interface. On the left, there's a sidebar with a search bar and a list of frequently used operations: Calculate, Convert column type, Filter, Math, Remove, Rename, Sort ascending, Sort descending, and Substitute. The main area is titled 'Operation' and contains a table with three columns: BOROUGH, MODEL, and SIZE. The table has 10 rows of data. The right side shows a 'Steps' panel with one step: 'Data Source' (Travel Rentalcar) and 'Join' (JUST ADDED). The 'Join' step is described as 'left-joined data from Travel Hotel based on columns BOROUGH,BOROUGH'.

- Click the **Operation** button to view the shaping operations menu.

- Click the **Rename** operation.
- Then **select column name** as **MODEL** > Click **Next** > **Rename column name with CARMODEL** > Click on **Apply**.
- Click the **Save** button on the toolbar to save the Data Flow.

Similar way you can rename following columns:

- **SIZE** with **CARSIZE**
- **RENT** with **CURRENT**
- **AGENCY** with **CARAGENCY**
- **PRICE** with **HOTELPRICE**
- **NAME** with **HOTELNAME**

13.5 Create a New Column

In the travel dataset there are **RENT** column represents daily rental cost of a car, where as **PRICE** represent daily hotel room rate. You need to create new column named **COST** that stores value of **RENT + PRICE**.

The screenshot shows the 'Operation' step in the Data Flow interface. The left pane displays the 'Calculate' operation configuration. It has selected the column 'CURRENT' (Integer type) and chosen 'Addition' as the calculation type. The 'HOTELPRICE' column is specified as the second operand. A checkbox for 'Create new column for results' is checked, and the resulting column is named 'COST'. The right pane, titled 'Steps', lists the following operations:

- Rename column (Renamed column SIZE to CARSIZE)
- Rename column (Renamed column RENT to CURRENT)
- Rename column (Renamed column AGENCY to CARAGENCY)
- Rename column (Renamed column PRICE to HOTELPRICE)
- Rename column (JUST ADDED)

SOURCE FILE: Travel Rentalcar SAMPLE SIZE: First 28138 rows

- Click the **Operation** button to view the shaping operations menu.
- Scroll down and click the **Calculate** operation.
- Select column **CURRENT**
- Click **Next**
- Choose **Addition** for Perform a calculation...
- Specify **Column** radio button
- Select column **HOTELPRICE**
- Check the box for Create new column for results.
- Use **COST** as the name of the new column for result

- Click **Apply**
- Click the **Save** button on the toolbar to save the Data Flow.

13.6 Run Data Flow

In order to process the shaping operations, you need to create a **Job** and run it. The job will use the data flow's output data set name, target location and format type to place and create the data flow output. Based on the changes you specified, the job will create a CSV file named **Travel Shaped** in your **Travel** project.

- Click the **Jobs** button on the toolbar.
- Select the **Save and create a job** menu item.

- Enter a Job Name of **Travel Shaped<id>** with the proper case, and spaces between the words.
- Use description as **Prepare the travel data for analytics**.
- Use the **Default Data Refinery XS** runtime, it should be pre-selected.
- Click the **Create and Run** button.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below the navigation bar, the 'Assets' tab is selected in the 'Travel' project. A search bar at the top of the main content area contains the placeholder 'What assets are you looking for?'. Below it, there are two sections: 'Data assets' and 'Data Refinery flows'. The 'Data assets' section has a table with columns: NAME, TYPE, CREATED BY, LAST MODIFIED, and ACTIONS. It lists four items: 'Travel Shaped.csv' (Data Asset, created by admin on 14 Apr 2020), 'Travel Rentalcar' (Data Asset, created by admin on 13 Apr 2020), 'Travel Hotel' (Data Asset, created by admin on 13 Apr 2020), and 'Db2 Travel' (Connection, created by admin on 13 Apr 2020). There are also '+' buttons to add new data assets or refinery flows.

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
Travel Shaped.csv	Data Asset	admin	14 Apr 2020, 5:21:28 pm	
Travel Rentalcar	Data Asset	admin	13 Apr 2020, 10:35:15 am	
Travel Hotel	Data Asset	admin	13 Apr 2020, 10:35:15 am	
Db2 Travel	Connection	admin	13 Apr 2020, 10:35:14 am	

- Click on the **Travel** project navigation link on the toolbar to get back to the sections of the project.
- Click on the **Assets** tab to view the project assets.

14 Congratulation!

You have created the new data asset named **Travel Shaped**. This is the CSV dataset the Data Refinery generated based on your data shaping operation. As a data engineer you can handed it over to the data scientist for their data science project.