

IBM Cloud Pak for Business Automation Demos and Labs 2022

Capture

IBM Automation Document Processing
V22.0.2

Lab Automation Document Processing

V 2.0

Clandis Baker
SWAT Business Automation Portfolio Specialist – Capture Products
bakercl@us.ibm.com

Krish Lakshminarayanan
Global Technical Program Leader for Capture / Intelligent Document Processing Global Sales (WW)
krishkrish@ibm.com

Ryan Sparks
Advisory Business Automation Tech Sales Leader – RPA/ADP
rmsparks@us.ibm.com

NOTICES

This information was developed for products and services offered in the USA.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk. IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

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 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.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

TRADEMARKS

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., 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.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

ITIL is a Registered Trade Mark of AXELOS Limited.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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.

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

© Copyright International Business Machines Corporation 2020.

This document may not be reproduced in whole or in part without the prior written permission of IBM.

US Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Table of Contents

1. Overview	6
1.1 Getting HELP during the lab	6
1.2 Abstract	6
1.3 Introduction.....	6
2 Getting started	8
2.1 IBM TechZone – Reserve the environment.....	8
2.1.1 Credentials.....	9
2.2 Set up WireGuard VPN	11
2.3 Open your IBM Cloud Environment	13
3 Lab Overview.....	16
3.1 How does ADP work?	16
4 Create Document Processing Project	18
4.1 Reviewing the interface.....	23
4.1.1 Build Tab	23
4.1.2 Enrich Tab	24
4.1.3 Configure Tab.....	25
5 Configure a Wage and Tax document type.	28
5.1 Create Wage and Tax document type.....	28
5.2 Create Field	30
5.3 Create the Employee Name Address field.....	32
5.4 Create Employee Social Security Number Field.....	33
6 Document Types and Samples Overview.....	37
6.1 Categorize documents.	38
7 Train classification.....	45
7.1 How do I improve my results?	49
8 Data extraction	51
8.1 Correcting extracted values	53
8.2 Train extraction model.....	58
9 Data standardization	59
10 Version and deploy your project	61
11 Application designer	63
11.1 Create your Runtime Application.	63
11.2 Upload documents for processing	68
11.3 Correct any classification errors.....	70
11.4 Correct extraction issues.....	72
12 Export Import Project.	80
Appendix A - Troubleshooting	82
TechZone Pending Status taking Long Time	82
Can't find user/password in Daffy	82
APPLICATION BLANK	84
Connection issue with Workstation to Cloud.....	85
OPENING AN INCOGNITO WINDOW.....	85

Appendix B - BAW & ADP Integration Sample	87
https://github.com/IBM/baw-adp-integration-sample	87

1. Overview

1.1 Getting HELP during the lab

- Slack channel on #cp4ba-tech-jam-capture.
- For internal IBM, another good resources the Archive slack channel for questions: #cp4ba-adp-lab or <https://ibm-cloud.slack.com/archives/C01LVVBMWPN>
- For external participants besides the Slack channel, use the Webex chat if you are in a webex event or just speak up
- For others, email bakercl@us.ibm.com. This method will be slower and will be best effort. It may require jumping on a Webex meeting to provide help.
- Getting help after lab reach out to the following:
 - bakercl@us.ibm.com
 - krishkirsh@us.ibm.com
 - rmsparks@us.ibm.com

1.2 Abstract

Set up a capture solution in minutes. Introduce technical sellers to IBM Automation Document Processing. In this session, students will configure their own capture project. They will learn how to use machine learning classification for their sample documents, define fields for extraction, create validation rules, and use deep learning* (subject to environment configuration) to automate data extraction.

1.3 Introduction

Welcome to the Automation Document Processing lab. This lab will introduce you to Document Processing and provide you with an understanding how you can configure it for your customer opportunities.

Automation Document Processing provides a tailored solution that reads your documents (in English, French, Spanish, German, Dutch, Portuguese), extracts data, and refines and stores the data for use.

With the right business knowledge, you can design deep learning models without being a data scientist. The Document Processing Designer includes pre-trained deep learning models that you can use as a base for your own model. The pre-trained document types include bills of lading, invoices, and utility bills.

You can extract text, check boxes, forms, tables, barcodes, signature detection and even free text. With no or low code options, you can create an application that processes documents, extracts data, flags issues, and stores your documents and data. And the data enrichment capabilities ensure that the extracted data is standardized and ready for use in downstream integrations.

This lab will not cover all the available functionality available due to time constraints. Additional labs will be created in the next few months to add to your knowledge and understanding of Document Processing.

2 Getting started

Download the sample documents in the zip file. You can find them here:

<https://github.com/IBM/cp4ba-labs/tree/main/22.0.2/Document%20Processing>

The screenshot shows a GitHub repository page for 'cp4ba-labs / 22.0.2 / Document Processing'. The 'Code' tab is selected. A red box highlights the file 'Design Forms Group1.zip' in the list of files. Other files listed include 'Group 2 ADP Application.zip', 'Readme.md', and '[In Progress]Lab Guide - Automation Document Proce...'. The interface includes standard GitHub navigation and commit history.

_1. Click on “Design forms Group1.zip”.

_2. Then Click on Download

The screenshot shows a GitHub file page for 'Design Forms Group1.zip'. The file size is listed as 1.23 MB. A red box highlights the 'Download' button at the bottom right of the file preview area. The interface includes standard GitHub navigation and commit history.

_3. Repeat above steps “Group 2 ADP Application.zip”

You will notice the images are in various folders that will be referenced in the lab later on.

2.1 IBM TechZone – Reserve the environment.

What is IBM TechZone?

IBM Technology Zone (techzone.ibm.com) enables IBM teams and IBM Business Partners to provision technical “Show Me” live environments, Proof-of-Technologies, prototypes, and

Minimum Viable Prototypes, which can be customized, shared with peers and clients to experience IBM Technology.

Learn more: <https://techzone.ibm.com/collection/onboarding#tab-1>

2.1.1 Credentials

- _1. Navigate to <https://techzone.ibm.com/collection/63457fcba311ed0018ca2442>

The screenshot shows a resource page titled "Pak Installer: Automated environments for OpenShift, IBM Cloud Paks, and Cartridges". The page includes a rating of 4 stars from 40 reviews, a "Build. Show. Share." icon, and a "Gold" badge. Below the title, there's a brief description: "The DAFFY automated Openshift + Cloud Pak installer has arrived to Techzone, fresh with a new name, **Pak Installer**. This asset was designed to help pre-sales (Tech Sales/BPs) with POC / MVP installs. This team has now enabled it within Techzone to quickly install Openshift + the latest/greatest Cloud Paks (CPD, CPI, CPBA) where Tech Sellers and Business Partners can stand up the entire stack in TechZone within hours. This Collection and Tiles use the same process as was with DAFFY, but through a simple 1 page input when you create a reservation, allowing users with any skill level to provision/use OpenShift and Cloud Paks to quickly get an environment and showing the value of the Pak." The sidebar on the left lists "Authors", "Resources", and "Comments".

Note: This environment is built with Daffy by Kyle Dawson with the latest releases. This environment can also be used at a customer site with same tool and framework of Daffy.

- _2. Click Cloud Pak for Business Automation tab and scroll down to the “Cloud Pak for Business Automation 22.01/22.0.2 – VMWare tile.
- _3. Click on Reserve
- _4. On Create a reservation screen select option for when to start

The screenshot shows a "Create a reservation" page for the "Cloud Pak for Business Automation 22.01/22.0.2 – VMWare" tile. The page has tabs: "Select a environment/infrastructure", "Select a reservation type", "Fill out your reservation", and "Complete". Under "Single environment reservation options:", there are two radio buttons: "Reserve now" and "Schedule for later". At the bottom are "Cancel", "Reset", and "Submit" buttons. To the right of the form is a cartoon illustration of a city skyline with a lightbulb and a plane.

- _5. Create a Reservation

Based on the reservation type you are making, provide the required information

Customer Demo : Need a short customer-facing demonstration

Practice/Self-Education: Need to gain experience

Standard proof of concept; Need an environment for a standard product use case.

Custom Proof of concept: Need a complex, customized environment.

Testing: Need to test a specific function, configuration, or customization.

- _6. For this lab **Testing** will give you 3 days plus the option to extend it for another week.
- Otherwise, you will need a legitimate opportunity to leverage another reservation type.
- _7. For Preferred Geography (required) select your preferred data center location

Preferred Geography (required)

Choose a preferred geography

AMERICAS - us-east region - wdc04 datacenter
AMERICAS - us-south region - dal12 datacenter

- _8. For VPN Access **choose Enable**

VPN Access (required)

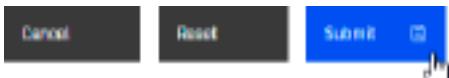
✓ Disable
Enable

- _9. For Starter Service **choose docprocessing**

Starter Service (required)

✓ all
content
content-decisions
decisions
docprocessing
workflow

- _10. Click Submit

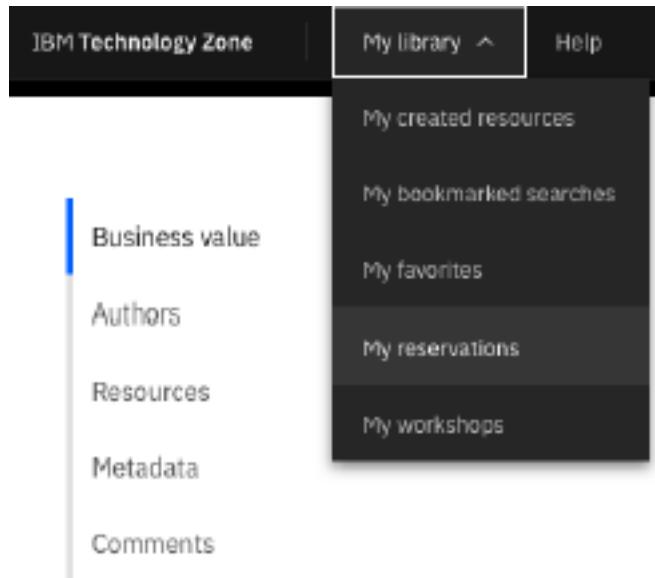


Upon receiving the Your environment is ready email, please allow up to 1 hour for the start-up services to fully complete. If after receiving email and a few hours have passed and your environment is not up, check [Appendix A – Trouble Shooting](#) for possible fix. Once the start-up process is complete you can click on the links identified in the email. However, it is recommended that you review your reservation information from the IBM Technology Zone – My reservation site.

- _11. Click My reservations



- _12. Once you get the email from the IBM Technology Zone site, you can access your environment reservation(s) by **clicking on the My library then My Reservations**

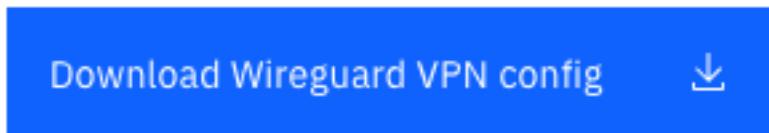


You can also access directly using the link below

<https://techzone.ibm.com/my/reservations>

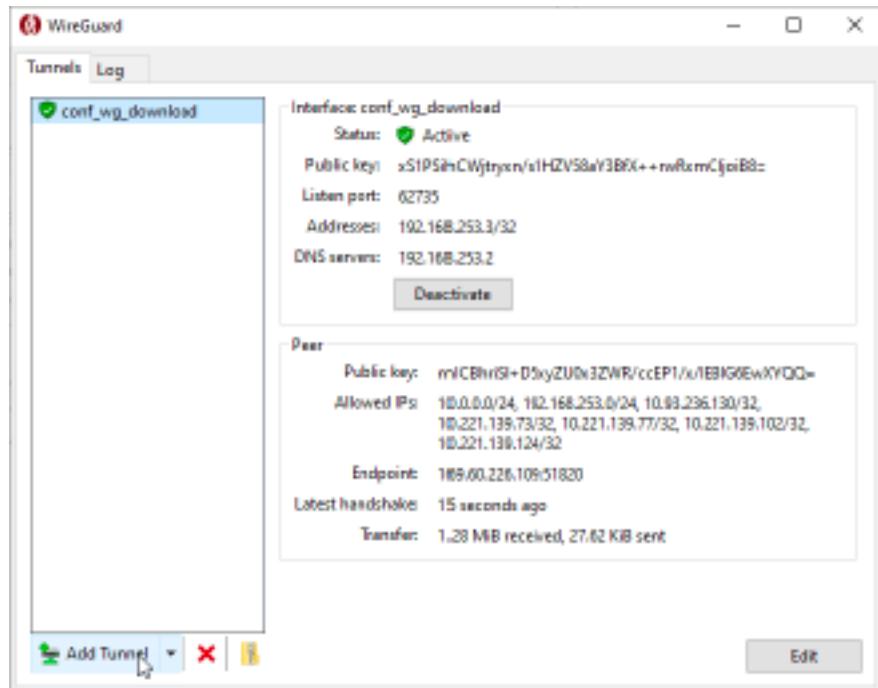
2.2 Set up WireGuard VPN

- _4. Open your reservation tile and scroll to bottom.
- _5. Click **Download WireGuard VPN config** button to download conf_wg_download.conf to your local workstation

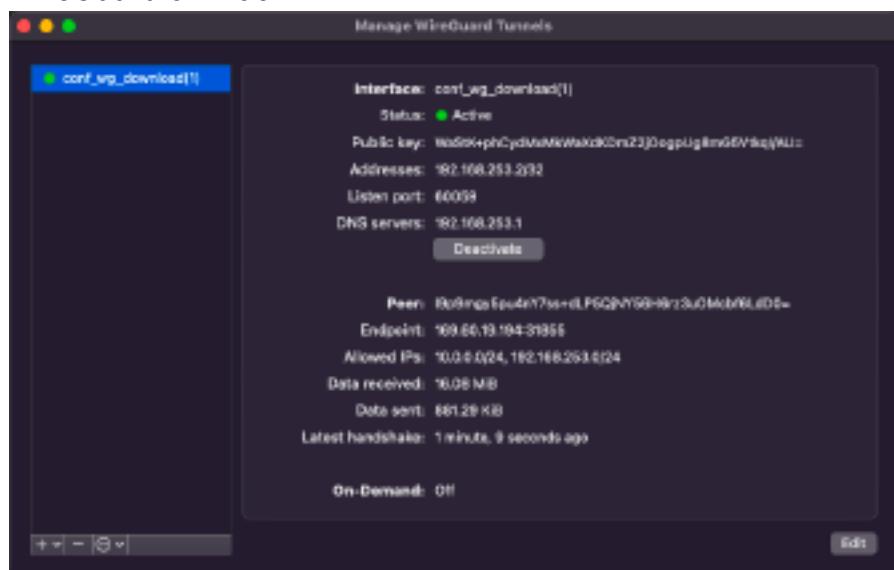


- _6. On your local workstation, install WireGuard by accessing <https://www.wireguard.com/install/>
- _7. Launch WireGuard
- _8. Click Add Tunnel and load the **conf_wg_download.conf** file.

WireGuard on Microsoft Windows



WireGuard on Mac



2.3 Open your IBM Cloud Environment

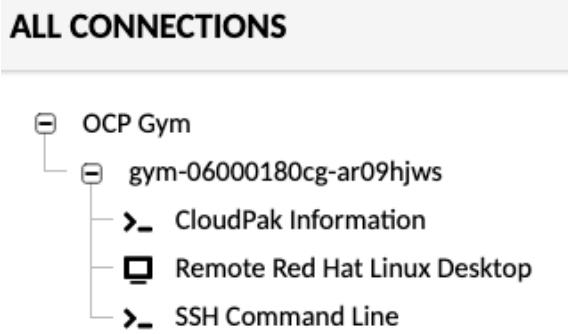
- _1. Back on your reservation screen **Click on Open your IBM Cloud environment**

The screenshot shows a reservation for 'Cloud Pak for Business Automation 22.0.1/22.0.2 - VMWare (Powered by Pak Installer)'. The reservation details include:

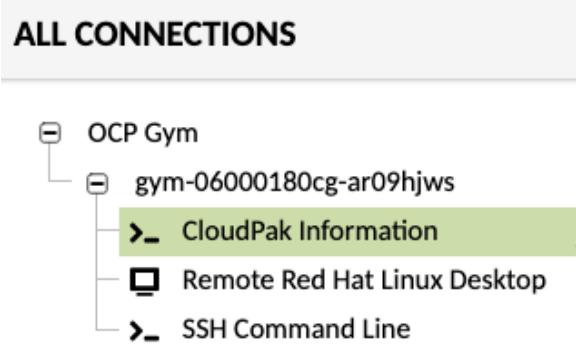
- Date: Feb 20, 2023 7:49 AM to Feb 27, 2023 7:49 AM
- Expires in: 2 days, 22 hours, 41 minutes
- Extend limit: 2
- Status: Ready

The 'Open your IBM Cloud environment' button is highlighted in blue.

- _2. **Expand OCP Gym under All Connections**



- _3. **Select CloudPak Information**



- _4. This will open Daffy Options window. **Enter 2** for Services

```
#####
#          Daffy Options          #
#####
Please use this tool and select what option you would like to retrieve more info on.
With this menu you can get your OpenShift Console URL, id/passwords and status.
You can also get your CP4BA Console URL, id/passwords and status info.

1) OpenShift
2) Services
3) ExitMenu
#? 2
```

- _5. **Enter 1** for Console information

```
#####
#          Daffy Options          #
#####
Please use this tool and select what option you would like to retrieve more info on.
With this menu you can get your OpenShift Console URL, id/passwords and status.
You can also get your CP4BA Console URL, id/passwords and status info.

1) OpenShift
2) Services
3) ExitMenu
#? 2
CP4BA Services Menu:
1) Console
2) Status
3) Back
#? 1
```

- _6. **Locate Username and Password** and copy and paste these to notepad. You will need to login into your environment.

Note: Controls for copy and paste in guacamole.

For Mac users:

CONTROL_OPTION_SHIFT

For Windows users:

CTRL_ALT_SHIFT

- _7. Back on your Reservation tile **copy** the **link Cloud Pak Dashboard URL** to your favorite browser.

Cloud Pak Dashboard URL

<https://cpd-cp4ba-starter.apps.ocpinstall.gym.lan>

You may get a Your connection is not private, if click advance then click Proceed.
This may occur twice.

_8. Login with user/password from step 6 above.

3 Lab Overview

The lab will focus on the design time tasks for Automation Document Processing (ADP). Despite the push for the digitization of content for many years, there are still a lot of paper documents that require workers to read and interpret the information – whether it is structured data, such as tax forms, or semi-structured data, such as invoices, utility bills, and so on. This lab describes how to set up and automated document processing pipeline using ADP.

3.1 How does ADP work?

Document Processing Designer

You use the Designer interface to create a set of document types and related fields that comprise your Document Processing project. Document Processing Designer combines an intuitive interface with a set of AI and deep learning tools that identify and learn the document types that matter to your organization. For each document type, you designate which pieces of information to extract as data for that document to be used by downstream applications. You can also apply tools to clean up and standardize the data as it is extracted.

Deployment tools

After you build the Document Processing project in the Designer, you deploy the project to make it available for building your document processing application. The deployment process is also used to configure the repository to receive the processed documents from your end-user application.

Application templates and toolkits

You use the no- or low-code application building capabilities of Application Designer, customized templates and toolkits, and the AI model of your Document Processing project to create a document processing end-user application. This application recognizes your documents, extracts your relevant data, and presents issues to fix before sending the documents to storage and using the data in other systems.

Document processing application and document management

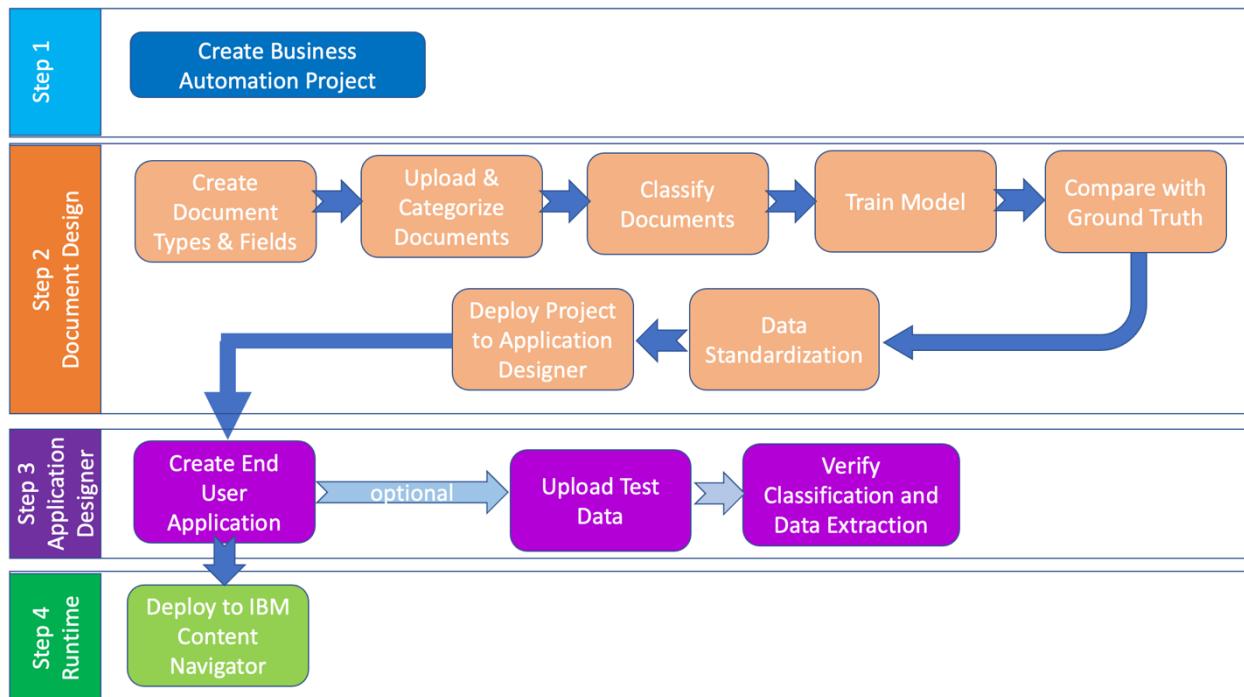
The application that you build uses the AI and deep learning to automatically detect, extract, and standardize the data in all your documents. Any anomalies are flagged according to your customized model and the priority that you set so that your document processing user can correct issues before the documents are finalized.

When you deploy your document processing application, you connect it to a content repository that manages the document types and the extracted data for each document. The solution is fully integrated with IBM FileNet® Content Manager, simplifying document and data storage by applying your existing filing architecture and business rules to each processed document. The content and metadata are automatically saved in FileNet within the appropriate document class.

End result

Your document types are stored in the content repository, with appropriate retention and access controls. An associated JSON file reflects all the extracted data for the document. Properties are set on the document with the data definition-controlled values. Your extracted data is cleaned, standardized, and ready for use in other applications.

The following diagram shows the tasks required to configure and deploy a new ADP project.



Step 1 – Create Business Automation Project

Each document processing project requires a separate repository in your Git organization. Coordinate with your Git administrator to create the repository for your project.

Step 2 – Document Design

This step shows the high-level tasks that will be needed to complete to train the system to recognize document types, successfully extract fields and tables, configure the fields in FileNet and finally deploying your content project to the application designer so you can configure the end-user interfaces.

Step 3 – Application Designer

The application designer is where you would configure end-user interfaces such as the classification and verification screens. The lab will not go in a lot of details on how to configure the interfaces. It will instead show you how to create an application, and test processing a batch of documents through the system

Step 4 – Runtime

End-users would be using the runtime IBM Content Navigator interface to process documents or batches, classify document and verify extracted field data in the verification screen.

4 Create Document Processing Project

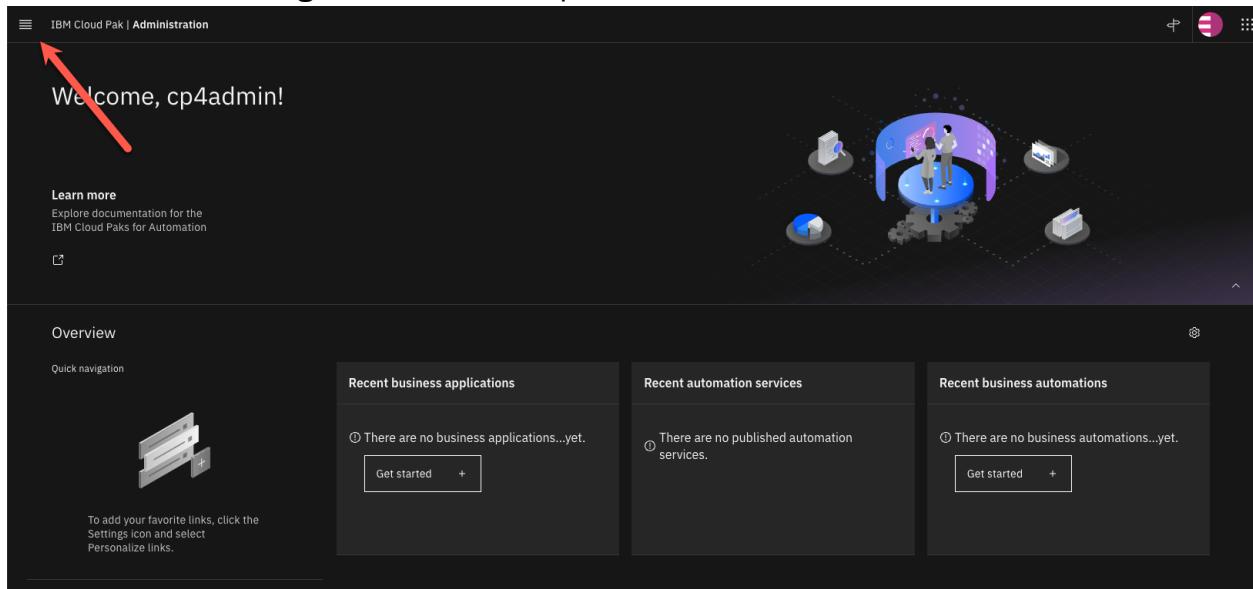
Step 1

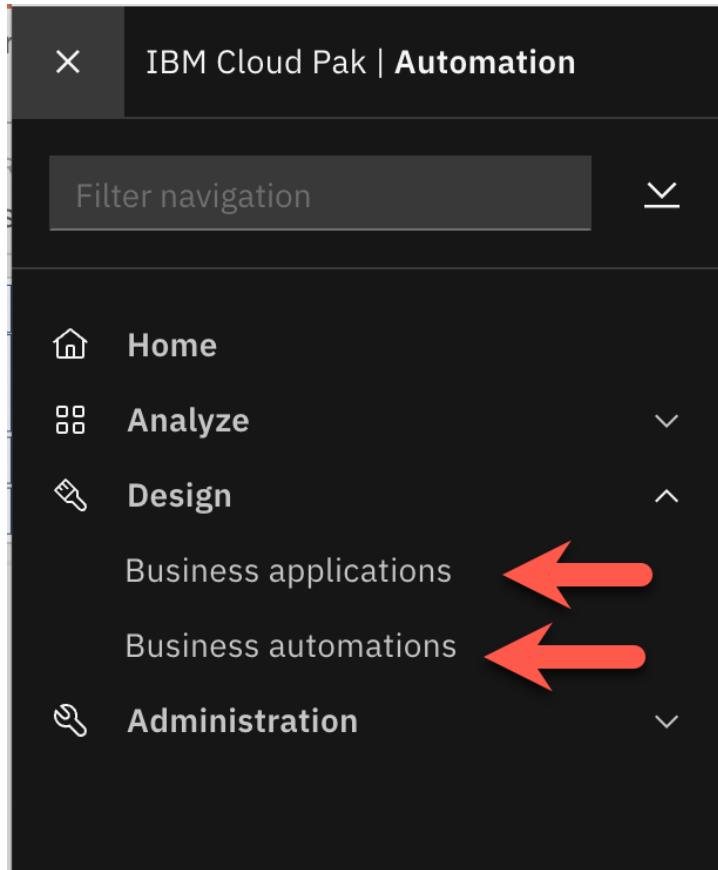
Create Business Automation Project

IBM Business Automation Studio is the single authoring and development environment for the IBM Cloud Pak for Automation platform that accelerates digital transformation. Business Automation Studio provides an entry point to various designers to help you reach your goals.

There are two distinct parts to the Business Automation Studio configuration.

_1. Click on the hamburger menu at the top left next to IBM Automation.





Business Automations provides the Document Processing configuration of the document classes, and the **Business Applications** provides the user interfaces.

Within the Business Automations you can create or reuse automations. An automation is a collection of artifacts that fulfills a business purpose. You can publish some automation artifacts as automation services that you can call and reuse in a consistent way. Also in Business Automation, you use the **Document Designer** interface within Automations to create a set of document types and related fields that comprise your Document Processing project.

The Document Processing Designer combines an intuitive interface with a set of AI and deep learning tools that identify and learn the document types that matter to an organization. For each document type, you designate which pieces of information to extract as data for that document to be used by downstream applications. You can also apply tools to clean up and standardize the data as it is extracted.

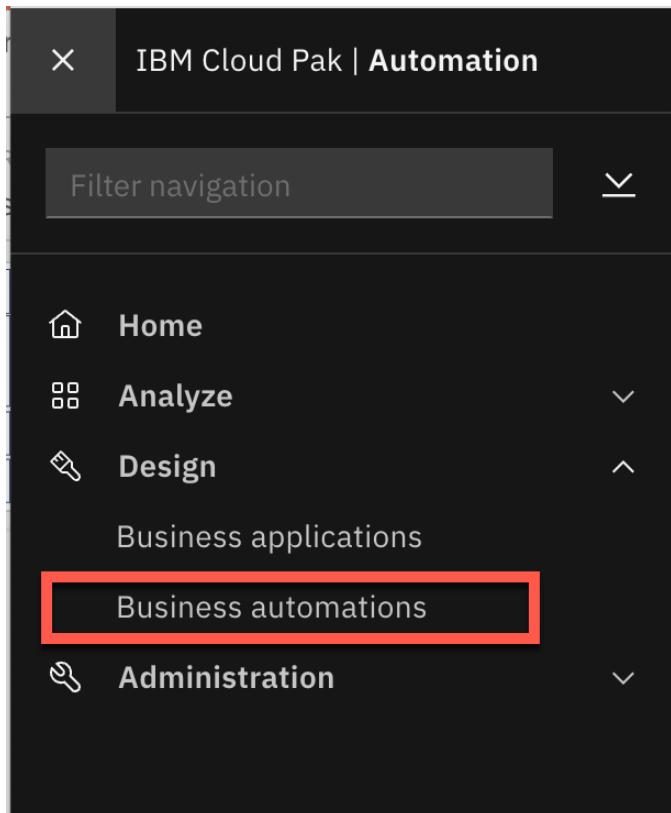
Within **Business Applications** you can quickly create user interfaces that integrate tasks, data, and automations. You can start with a template to ensure consistency. You can also use toolkits to share artifacts from existing applications.

We will start with the Business Automations.

Once logged in to the IBM Automation Server, you should see the Welcome screen.



_2. Click on **Drop down arrow** next to Design then **Select Business Automations**.



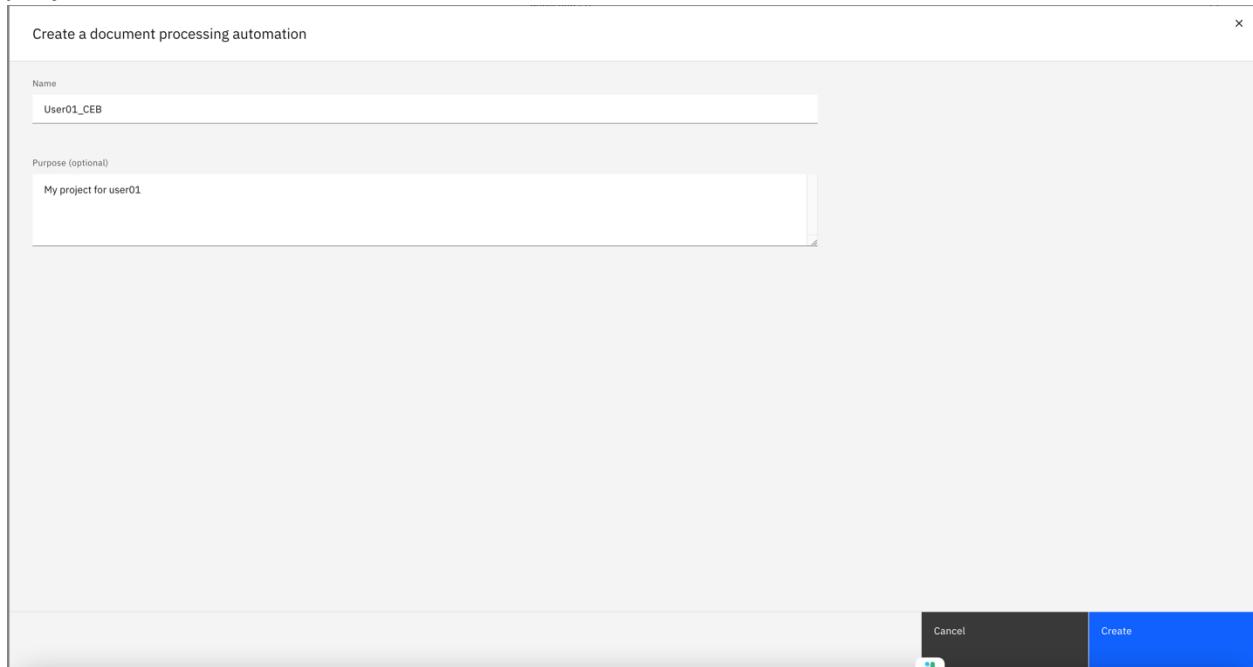
You may be presented with an overview screen. **Select Maybe Later**. Then following screen appears.

The screenshot shows the 'Business automations' page in the IBM Cloud Pak | Automation interface. At the top, there is a navigation bar with a menu icon and the text 'IBM Cloud Pak | Automation'. Below the header, the title 'Business automations' is displayed. A brief description follows: 'Create or reuse automations. An automation is a collection of artifacts that fulfills a business purpose. You can publish some automation artifacts as automation services that you can call and reuse in a consistent way.' A 'Learn more' link is provided. Below the description is a control bar with 'Create' (blue button), 'Import' (dark grey button), and a downward arrow. A list of automation types is shown, each with a right-pointing arrow: 'Published automation services', 'Decision', 'Document processing', 'Workflow', and 'External'.

_9. Click on the **Create** twisty and select **Document processing automations**.

The screenshot shows the 'Business automations' page again, but now the 'Create' button is highlighted with a red arrow pointing to it. A dropdown menu has opened, listing several automation categories. The category 'Document processing automations' is highlighted with a red box. Other visible items in the dropdown include 'Decision automations', 'Workflow', and 'External'. The main list below the dropdown includes 'Document processing', 'Workflow', and 'External', each with a right-pointing arrow.

_10.In the Create a document processing automation window **enter a name** for the project.



_11. Click on **Create** in the lower right-hand corner.

4.1 Reviewing the interface.

The screenshot shows the IBM Cloud Pak | Automation interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a user profile icon. Below it, the title 'User01_CEB' is displayed. The main area has three tabs: 'Build' (selected), 'Enrich', and 'Configure'. The 'Build' tab contains five sections: 'Document types and samples' (3 types, 26 samples on average, Ready), 'Classification model' (3 types trained, 100% accuracy, Ready), 'Extraction model' (3 types trained, 95% accuracy, Ready), 'Data standardization' (Not ready, Start button), and 'Document retention' (3 types reviewed, Ready). Top right are 'Share' and 'Version / Deploy' buttons. Below the tabs, there are 'Last shared | 2 minutes ago' and 'Latest version | not yet Deployed | not yet' status indicators.

Upon opening the project, there are three major sections: **Build tab**, **Enrich tab**, and **Configure tab**.

On the top right, you find the SHARE and VERSION/ DEPLOY buttons.



The SHARE button is used to save your configuration to your GitHub repository.

The VERSION / DEPLOY button is used to create a snapshot, or version of your configuration. Like the SHARE button, the VERSION button will save your configuration, but will also create a version of it while retaining your previous version.

Once you have created a version of your configuration, you can also use this button to DEPLOY your version to the Business Applications area of ADP. You need to do this before you can go into the Business Application tile and configure your user interfaces.

4.1.1 Build Tab

This is what we will be spending most of our time on. The BUILD tab shows the guided configuration for building a Document Processing project. It shows the five steps required.

Document types and samples: Here we will define the document types that can be recognized by this automation and upload sample documents for training. By default, any project will be pre-populated with three pre-trained document types (Bill of Lading, Invoice, and Utility Bill).

Classification model: classification: Here we will teach the system how to recognize the different document types.

Extraction model: Here we will teach the system how to extract information for each document type based on the classification.

Data Standardization: This allows further refinement of the extracted information. For example, we want to standardize all dates to be formatted as YYYY/MM/DD. Having a standardized data format will help with any subsequent automation process.

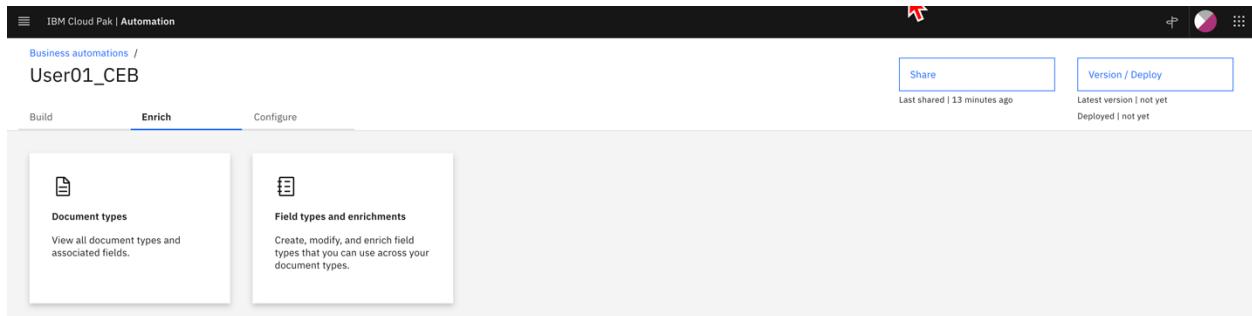
Document retention: This allows us to define how long we want our documents to be kept in the system. Documents that have exceeded the retention period will be automatically expunged. This could be important for regulatory compliance or for managing the overall storage size.

Section	Status	Count	Accuracy	Action
Document types and samples	Ready	3 types	26 samples on average	Open →
Classification model	Ready	3 types trained	100% accuracy	Open →
Extraction model	Ready	3 types trained	95% accuracy	Open →
Data standardization	Not ready			Start →
Document retention	Ready	3 types reviewed		Open →

4.1.2 Enrich Tab

_1. Click on the ENRICH tab.

Enrich provides a quick way to define your document types and the fields you wish to extract. In this section, we can define additional enrich rules. An example of an enrich rule is to specify the expected format for an invoice number (all numerical) or a driver's license. The more we can tell document processing about how different data will be formatted, the higher the chance it will recognize the information.



_2. Click on **FIELD TYPES AND ENRICHMENTS** to begin. In this tile, you will see some of the pre-configured fields in the *SYSTEM LIBRARY*. Customers can use these fields in their document type field definitions as needed.

Field type	Value type
Address block	String
Address information	Composite
Addressee	String
Boolean	Boolean
Building number	String
City	String
Country	String
Country code	String
Country name	String
Currency	Composite
CurrencyCode Object Type	String
Date	Date
Decimal	Decimal
Email	String

_3. Click on <*your project name*> in the bread crumb trail at the top.

4.1.3 Configure Tab

_4. Click on **Configure Tab**

This is where we can configure other operational aspects of the project. The export project creates a .zip file that contains the document types, field types and enrichments, which you can use to start training with new sample files. You can also decide to include the training model and the sample training files in your export if you want to move your entire project to a new instance of Document Processing for example. To import a project, select the .zip file to import. When you import a .zip file you have two options: overwrite the existing project or merge the existing project. If you merge the existing project, document types, field types, enrichments, and sample training files are imported unless there is a conflict. Models are not imported.

The screenshot shows the 'Configure' tab selected in the navigation bar. On the left, there's a sidebar with 'Import / Export ontology', 'Language settings', and 'Git server configuration'. The main area has two main sections: 'Export project' (with a 'Export project' button) and 'Import project' (with a 'Import project' button). Top right: 'Share' (Last shared | a day ago) and 'Version / Deploy' (Latest version | v2 | a day ago, Deployed | v2 | a day ago).

In Extraction language, select which languages are used in the documents that you plan to process. You can choose English, Dutch, French, German, Brazilian Portuguese, or Spanish. Make sure to choose only the language or languages that are likely to be used in your document sets. Choosing more than one language can affect the accuracy of your document processing model.

In Display name language, select the language that you use to enter display names for fields and document types. These are the names that are displayed in the Designer and in the applications. The display name language is also used in the Content Engine as the localized string locale setting for document classes and properties. Document Processing project deployment supports only one language per project. If your organization has multiple projects with different language settings, these projects cannot be deployed to the same Content Engine server if they share common properties. For example, when you define data definitions during data standardization, you cannot map a field to an existing data definition that was created in a different language.

The screenshot shows the 'Configure' tab selected. In the sidebar, 'Language settings' is selected. The main area shows 'Language Settings' with 'Extraction language' set to English. It also shows 'Display name language' set to English (en) (default). Bottom right: 'Cancel' and 'Save' buttons.

The Git server configuration is where you create a connection to the Git server for the first project that you create in Document Processing Designer. This setting applies to all subsequent projects that you create.

IBM Cloud Pak | Administration

Business automations / Clandis Baker Project

Build Enrich Configure

Import / Export ontology

Language settings

Git server configuration

In order to share, version and deploy, you need to establish a connection to your organization's Git server.

Git vendor: Gitea

Git server organization URL: <https://cp4deploy-gitea-svc:3000/content-designer>

Git server REST API URL: <https://cp4deploy-gitea-svc:3000/api/v1>

Username: git

Type of credentials: API key Password

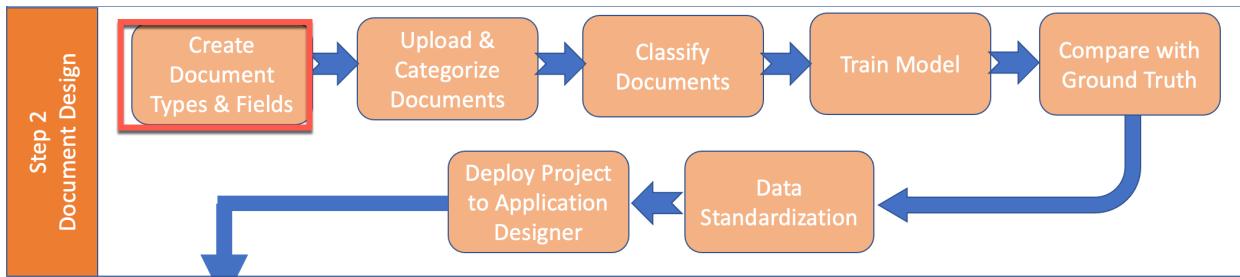
Credentials: Enter a password or API key

Share Last shared | a day ago

Version / Deploy Latest version | v2 | a day ago
Deployed | v2 | a day ago

Test Save

5 Configure a Wage and Tax document type.



Before we use the guided configuration, you will configure some additional document types and fields used to extract data prior to uploading sample documents.

To do this lab, we will use the ENRICH tab to add fields to a newly created Wage and Tax document type.

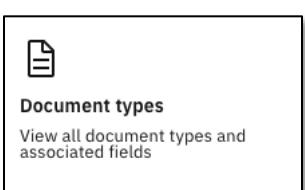
5.1 Create Wage and Tax document type.

- _1. Click on **<your project name>** in the breadcrumb trail to return to the start page. In the example below our project was called **<User01_CEB>** if not already on the Project page

The screenshot shows the 'IBM Cloud Pak | Automation' interface. The top navigation bar includes 'IBM Cloud Pak | Automation', a user profile icon, and a three-dot menu. Below the bar, the breadcrumb trail reads 'Business automations / User01_CEB'. A red arrow points to the project name 'User01_CEB'. The main content area is titled 'Field types and enrichments'. It contains a message about modifying field types and creating new ones. Below this are sections for 'Field type libraries' (with a 'sys' entry) and 'Field types'. On the right side, there's a 'Address block' section with a 'View' button. A blue box highlights the 'Import library' button in the top right corner.

- _2. Click on the **ENRICH** tab

- _3. Click on DOCUMENT TYPES



We will now create a document type for Wage and Tax documents and fields to extract data from them.

- _4. Click on the CREATE DOCUMENT TYPE button in the top right corner.



- _5. The Add document type window pops up. Enter Wage and Tax for the display name. There is no need to enter a symbolic name ADP will use the display name a

base. There's no need to add description in this lab unless you want to.

Add document type

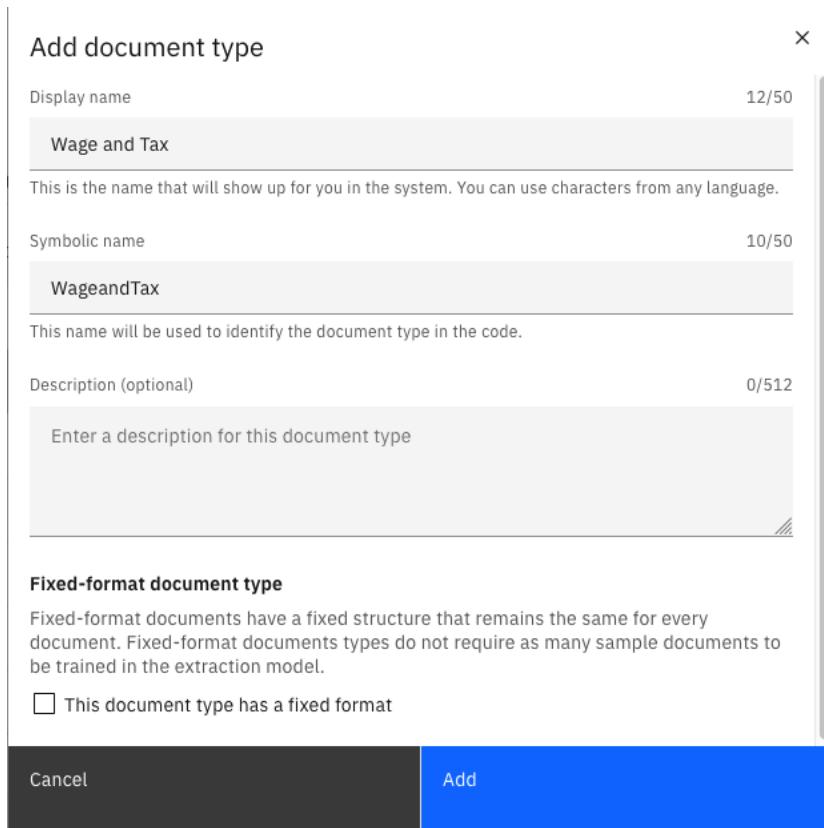
Display name 12/50
Wage and Tax
This is the name that will show up for you in the system. You can use characters from any language.

Symbolic name 10/50
WageandTax
This name will be used to identify the document type in the code.

Description (optional) 0/512
Enter a description for this document type

Fixed-format document type
Fixed-format documents have a fixed structure that remains the same for every document. Fixed-format documents types do not require as many sample documents to be trained in the extraction model.
 This document type has a fixed format

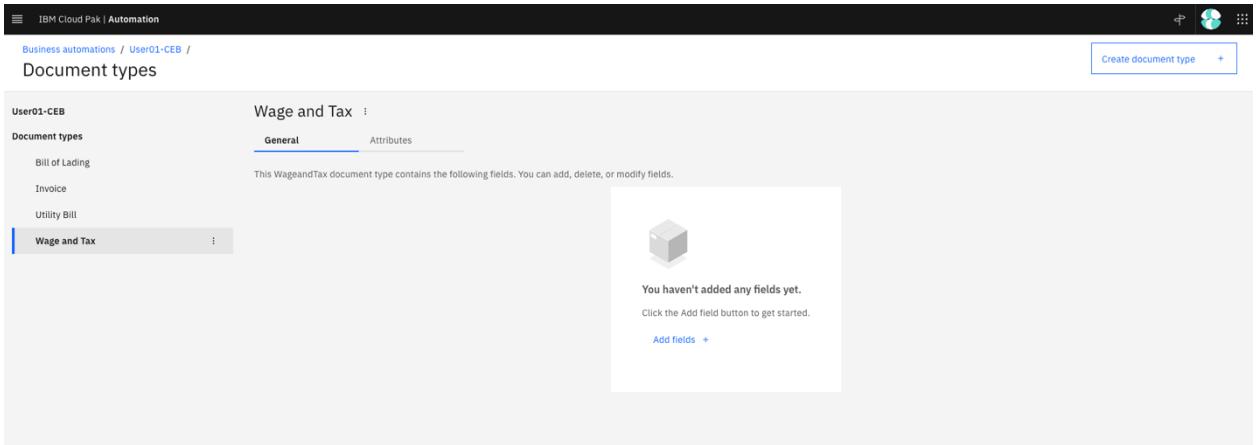
Cancel Add



Note: Notice the option for “Fixed-format document type”. If your form is static in nature or has a fixed structure that does not change, select this option so you will not have to provide as many samples. In our use case Wage and Tax documents have a variety of formats and are not static.

6. Click the ADD button.

You should now see your new document type (class) in the list of classes on the left.



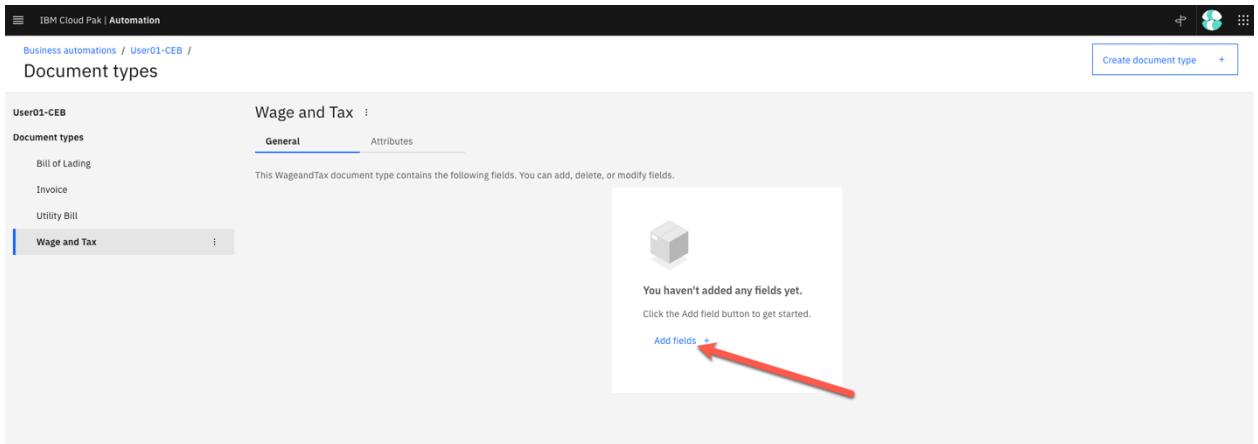
The screenshot shows the 'Document types' section of the IBM Cloud Pak | Automation interface. On the left, a sidebar lists 'User01-CEB' and 'Document types' with options like 'Bill of Lading', 'Invoice', 'Utility Bill', and 'Wage and Tax'. The 'Wage and Tax' option is selected and highlighted with a blue border. The main content area is titled 'Wage and Tax' and has tabs for 'General' and 'Attributes'. A message box in the center states 'You haven't added any fields yet.' and 'Click the Add field button to get started.', with a blue 'Add fields +' button. In the top right corner, there's a 'Create document type +' button.

_7. Select your Wage and Tax type. On the right, you should see an empty table of fields.

5.2 Create Field

We can now add some fields to the class.

_1. Click ADD FIELDS



The screenshot shows the 'Add fields' interface for the 'Wage and Tax' document type. It features a message box stating 'You haven't added any fields yet.' and 'Click the Add field button to get started.', with a blue 'Add fields +' button. A red arrow points to this 'Add fields +' button. The background shows the same 'Document types' interface as the previous screenshot.

_2. Enter the following values under the GENERAL Settings header

The screenshot shows the 'Create field' interface in IBM Cloud Pak | Automation. The document type is set to 'Purchase Orders'. The 'General' tab is selected. In the 'Display name' field, 'Ex. Employee's name, Le nom de l'employé' is entered, with a red error message 'This is a required field' displayed below it. The 'Symbolic name' field contains 'Enter a name'. The 'Field type' dropdown is set to 'sys:String'. Under 'Aliases', there is a text input field with '+'. The 'Description (optional)' and 'Value settings' tabs are also visible.

- **Field Name: Federal Income Tax Withheld**
- **Field Type:**
 - **Sys:Decimal**
- **Is this field required: Yes**
- In Aliases enter other possible names. Case and punctuation are very import when creating aliases. Enter the alias listed below. **Press the “+” after entering each one or press Enter key:**
 - **2 Federal income tax withheld**
 - **2. Federal income tax (note: the number two has a period after it)**

You should now see the following:

The screenshot shows the 'Create field' interface for 'Wage and Tax'. The document type is 'Wage and Tax'. The 'General' tab is selected. In the 'Display name' field, 'Federal Income Tax Withheld' is entered. The 'Symbolic name' field contains 'FederalIncomeTaxWithheld'. The 'Field type' dropdown is set to 'sys:Decimal'. Under 'Aliases', '2 Federal income tax withheld' and '2. Federal income tax' are listed. The 'Description (optional)' and 'Value settings' tabs are also visible.

_3. Click the NEXT button.

- _4. Click **NEXT** again on the Field patterns screen. You will not be adding patterns in this lab. Patterns are regular expressions that can be used as an alternative to aliases.

You should now be on the **VALUE SETTINGS** page. This is where you can set up validators, formatters, and converters.

- _5. Click **Create** your screen should look like this with your first field created.

Name	Type	Required	Sensitive
Federal Income Tax Withheld	Decimal	true	false

5.3 Create the Employee Name Address field.

- _1. Click **Add fields**.

Give it the following parameters:

- Field name: **Employee Name and Address**
- Field Type = **sys:Address information**
- Required = **yes**
- Enter the following other possible names (aliases):
 - ***Employee name and address***
 - ***e Employee's first name and initial Last name Suff***
 - ***e Employee's name, address, and ZIP code***
 - ***e/f Employee's name, address, and ZIP code***
 - ***e. Employee Name & Address***
 - ***e Employee's first name and initial***

By default, the system will use the field name as an alias. So, you do not have to add it.

For example, below, Employee Name and Address (field name), would be automatically used as an alias even if you do not add it to the list

The screenshot shows the 'IBM Cloud Pak | Automation' interface. In the top navigation bar, there are links for 'Business automations / DanO-CustTest-Keep / Document Types / Employee name and address'. On the right side, there are 'Cancel' and 'Next' buttons. The main area is titled 'Employee name and address' with a note 'Document type: Form W2'. Below this, there are tabs for 'General', 'Field patterns', 'Value settings', and 'Subfields', with 'General' selected. The 'General Settings' section contains fields for 'Display name' (set to 'Employee name and address'), 'Description (optional)' (with placeholder 'Enter a description for this field'), 'Symbolic name' (set to 'Employee name and address'), 'Field type' (set to 'sys:Address information'), 'Aliases' (with placeholder 'Enter an alternative name'), and checkboxes for 'This field is required' (checked) and 'This field contains sensitive information' (unchecked). A list of aliases is shown at the bottom: 'Employee name and address', 'Employee's first name and initial Last name Suff.', 'Employee's name, address, and ZIP code', 'Employee's name, address, and ZIP code', 'Employee Name & Address', and 'Employee's first name and initial'.

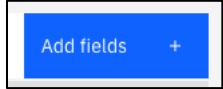
_2. Click **Next** no field patterns will be created.

_3. Click **Next** no value settings will be created.

_4. Click **Create** to finish creating the Employee Name and Address.

5.4 Create Employee Social Security Number Field

_1. Click on ADD FIELDS



Enter the following values in the GENERAL page.

- Field Name: **Employee Social Security Number**
- Field Type: **sys:Social Security Number**
- Is value required: **Yes**
- Other possible names (aliases). Remember, press RETURN on your keyboard between each entry:
 - **a Employee's social security number**
 - **a Employee's social security no.**
 - **a Employee's SSA number**
 - **a. Employee Social Security Number**
 - **Employee social security number**

Your screen should now look like the image below:

The screenshot shows the 'Employee Social Security Number' field configuration screen. At the top, there are tabs for 'General' (selected), 'Field patterns', and 'Value settings'. The 'General' tab has fields for 'Display name' (Employee Social Security Number) and 'Symbolic name' (EmployeeSocialSecurityNumber). Below these are sections for 'Field type' (sys:Numeric), 'Required' (checked), and 'Sensitive information' (unchecked). On the right, there are sections for 'Description (optional)' and 'Aliases'. The 'Aliases' section contains several entries: 'a Employee's social security number', 'a Employee's social security no.', 'a Employee's SSA number', 'Employee social security number', and 'a Employee Social Security Number'. At the bottom right are 'Cancel' and 'Next' buttons.

_2. Click NEXT

_3. Click NEXT again on the Field Patterns screen.

_4. Click Create on the Value settings.

_5. Create the following additional Fields.

The following table contains the values to use when adding the additional fields.

Follow the steps from the previous section to add the following fields

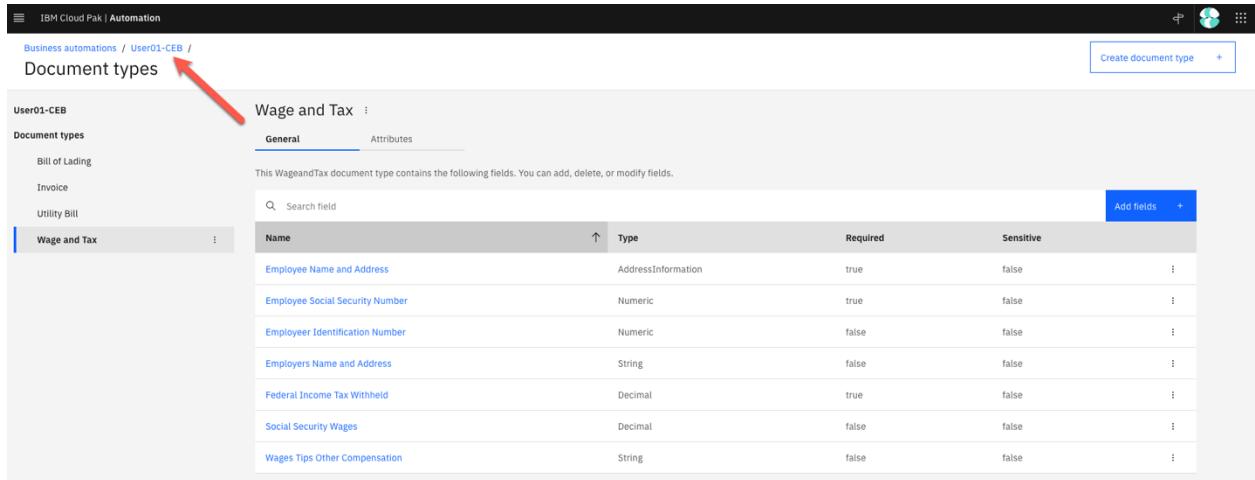
Field Name	Description	Type	Mandatory	Aliases
Employer Identification Number		sys:String	N	<ul style="list-style-type: none"> • b Employer identification number (EIN) • b Employer's FED ID number • b. Employer ID number • Employer identification number
Employers Name and Address		sys:String	N	<ul style="list-style-type: none"> • c Employer's name, address, and ZIP code • c Employer's Name & Address • Employers name and address
Social Security Wages		sys:Decimal	N	<ul style="list-style-type: none"> • Social security wages • 3 Social security wages
Wages Tips Other Compensation		Sys:Decimal	N	<ul style="list-style-type: none"> • 1 Wages, tips, other compensation • Wages, tips, other comp. • 1 Wages, tips, other comp. • 1. Wages tips, other comp • Wages tips other compensation

Reference for various field types:

Note: The basic default field types included in ADP are found here in the documentation

<https://www.ibm.com/docs/en/cloud-paks/cp-biz-automation/22.0.1?topic=enrichments-field-types-document-processing>

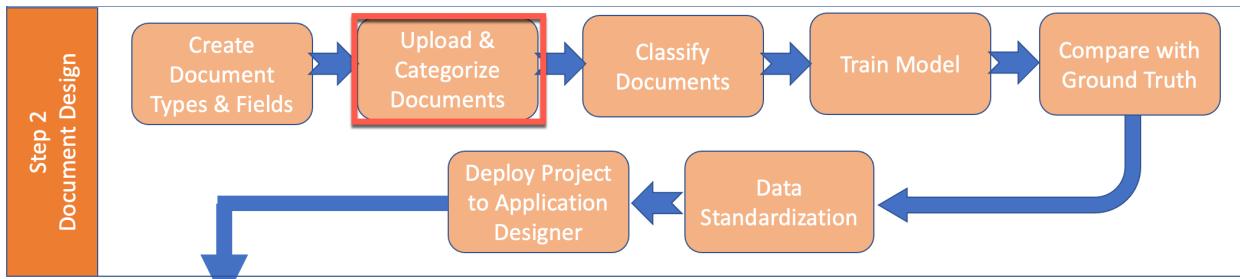
_6. Click on the <name of your project> in the breadcrumb link in the top left of your screen. In the following example the name of the project is <User01_CEB>.



The screenshot shows the 'Document types' page within the 'User01-CEB' project. The breadcrumb navigation bar at the top displays 'IBM Cloud Pak | Automation', 'Business automations / User01-CEB / Document types'. The main content area is titled 'Wage and Tax' under the 'General' tab. It contains a table of fields with columns: Name, Type, Required, and Sensitive. Fields listed include Employee Name and Address, Employee Social Security Number, Employee Identification Number, Employers Name and Address, Federal Income Tax Withheld, Social Security Wages, and Wages Tips Other Compensation.

Name	Type	Required	Sensitive
Employee Name and Address	AddressInformation	true	false
Employee Social Security Number	Numeric	true	false
Employee Identification Number	Numeric	false	false
Employers Name and Address	String	false	false
Federal Income Tax Withheld	Decimal	true	false
Social Security Wages	Decimal	false	false
Wages Tips Other Compensation	String	false	false

6 Document Types and Samples Overview



At this point in the process, we have created a new document type and configured the field names we want to extract off the document. For the system to know what to extract from your documents, it needs to be able to classify the documents. In this lab, we will teach the system to recognize the various document types on your system.

In the first part of the classification lab, you will explore the system's ability to automatically group similar documents together. This can be used to discover document types in a file share for example. You can also upload documents and have the system tell you what it finds. You would then use this information to create document types so you can classify the documents and data extract fields.

The project template comes pre-loaded with three document types: Bill of Lading, Invoice, and Utility Bill. In the last step we added a new document type Wages and Tax. In the BUILD tab of your project, you should now be seeing 4 document types. The three pre-loaded documents already have documents in them. You will be adding documents to the Wage and Tax document type. Your actual screen may vary from the following screen shot.

You will be asked to review the document categories the system finds and create the appropriate document types as needed.

Section	Status	Details
Document types and samples	Ready	4 types, 19 samples on average
Classification model	Ready	3 types trained, 100% accuracy
Extraction model	Ready	3 types trained, 95% accuracy
Data standardization	Not ready	
Document retention	Ready	4 types reviewed

6.1 Categorize documents.

For categorizing, we will have the system help us group similar documents together. To get started,

_1. Click anywhere in the Document types and samples box.

Category	Status	Count	Accuracy
Document types and samples	Ready	4 types	22 samples on average
Classification model	Ready	3 types trained	100% accuracy
Extraction model	Ready	3 types trained	97% accuracy
Data standardization	Not ready		

The CATEGORIZE feature analyzes each document and tries to find similarities between them. Based on these similarities, the system will divide the samples into categories for you to review. You can add documents or entire categories into either an existing document class or create new classes as needed.

Let's see what that looks like.

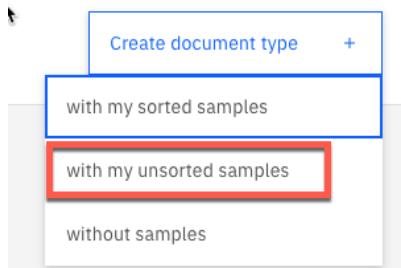
_2. Click on **CREATE DOCUMENT TYPE** in the top right of the screen.

- [Create document type +](#)
- [with my sorted samples](#)
- [with my unsorted samples](#)
- [without samples](#)

If you have the same document types already separated into folders, you can choose the first option, *with my sorted samples*. The system would simply ingest the documents from each folder into a different group.

For this exercise, we will select the second option, *with my unsorted samples* and let the system sort the documents for us. Use this option when you don't know how many different document types there are.

_3. Select the second option titled **with my unsorted samples.**



You should have already downloaded the files from [Section 3](#) to your laptop. You can either drag the folder to the window or select upload and grab all the files from where they were downloaded to on your laptop.

_4. Click Upload to get document samples.

From the downloaded sample documents open the folder name Design Forms Group1

Note: this will take several minutes (approximately 10 minutes), good time for some coffee or a stretch. Make sure to check ALL documents have been uploaded there are two pages or 12 items to verify.

_5. Click on the CATEGORIZE button.

The screenshot shows a web-based interface for document classification. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a user profile icon. Below it, a breadcrumb trail shows 'Business automations / User01-CEB / Document types and samples / Create document types'. On the right are 'Cancel' and 'Categorize' buttons. A search bar at the top left says 'Search sample documents' with an 'Upload' button. Below the search bar is a list of 12 PDF files under 'Document name': Mortgage Agreement1.pdf, Mortgage Agreement2.pdf, Mortgage Agreement3.pdf, Mortgage Agreement4.pdf, Mortgage Agreement5.pdf, TR_FW2_1001_0000_PS.pdf, TR_FW2_2000_0000_PS.pdf, TR_FW2_3000_0000_PS.pdf, TR_FW2_3001_0000_PS.pdf, TR_FW2_4000_0000_PS.pdf, UBILLCable_081_1_11.pdf, and UBILLCable_082_1_11.pdf. The interface includes a sidebar with 'Items per page' set to 20, and a footer showing '1 - 12 of 12 items'.

Note: The results may vary based on the documents uploaded, what the system already has learned, the version of ADP and more. Please look at this lab exercise from a high level. The categories you will be presented are the system's best guess on how they should be separated.

You will need to:

- Review the categories to see if the documents were separated correctly.
- Move documents into either a NEW document type or into an EXISTING document type.
- There should be 3 types in the samples you were provided.
 - Wage and Tax
 - Utility bills
 - Mortgage Agreements
- You will need to assign either an entire category (i.e., all sample documents) or individual documents in each category to the Wage and Tax and Utility bills document types which already exist on your system.
- You will need to create a new document type for Mortgage Agreements.

After a few seconds, the system will mark the documents with a status of ready as seen in the above image.

_6. Click on each of the categories to see what was grouped together as shown below.

NOTE: The names of the files are not used in any way in this process. The files were merely named this way to make it easier for you to quickly ascertain whether the documents were grouped correctly.

The screenshot shows the 'Create document types' page. On the left, there's a sidebar with 'Categories (3)' containing 'Category 1' (selected), 'Category 2', and 'Category 3'. Below this is a 'Document types (4)' section with 'Bill of Lading' (21 samples), 'Invoice' (31 samples), 'Utility Bill' (25 samples), and 'Wage and Tax' (0 samples). The main area shows 'Category 1 sample documents (2)'. It has a search bar 'Search sample documents' and an 'Upload' button. Two PDF files are listed: 'UBILLCable_081_1_1.1.pdf' and 'UBILLCable_082_1_1.1.pdf', each with a checkmark and a downward arrow icon.

The screenshot shows the 'Create document types' page. The sidebar shows 'Category 2' selected. The 'Document types (4)' section is identical to the first screenshot. The main area shows 'Category 2 sample documents (5)'. It has a search bar and an 'Upload' button. Five PDF files are listed: 'Mortgage Agreement1.pdf' through 'Mortgage Agreement5.pdf', each with a checkmark and a downward arrow icon.

The screenshot shows the 'Create document types' page. The sidebar shows 'Category 3' selected. The 'Document types (4)' section is identical to the first screenshot. The main area shows 'Category 3 sample documents (5)'. It has a search bar and an 'Upload' button. Five PDF files are listed: 'TR_FW2_1001_0000_P5.pdf' through 'TR_FW2_4000_0000_P5.pdf', each with a checkmark and a downward arrow icon.

At the time of writing this documentation ADP was able to categorize the sample set into each category. This is not always the case, sometimes document types will be combined into one category, so it's very important to look at each category and verify documents.

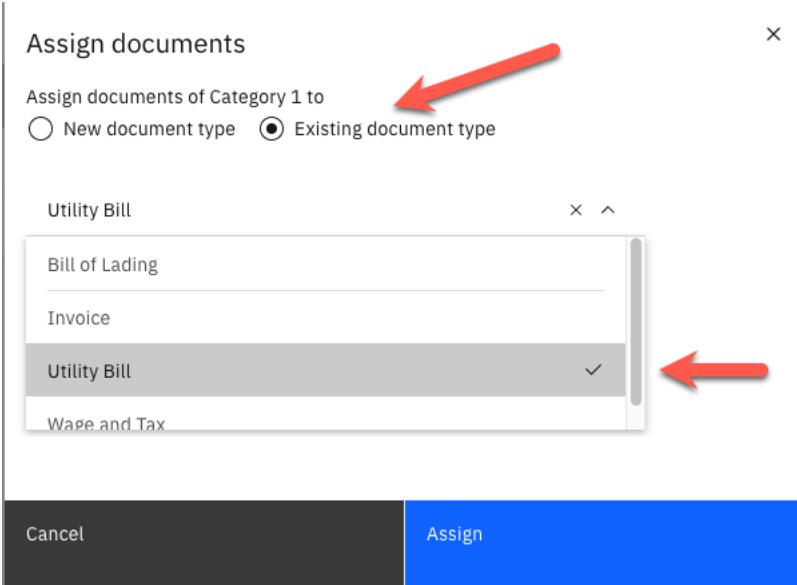
_7. If all documents within a category are correct as illustrated in the following screen shot, hover over the category name and **Click on the 3 dots** at the end of the category name.

The screenshot shows the 'Create document types' page in the IBM Cloud Pak | Automation interface. On the left, there's a sidebar with 'Categories (3)' and 'Document types (4)'. Under 'Categories', 'Category 1' is selected and has a context menu open, with 'Assign to document...' highlighted. The main area shows 'Category 1 sample documents (2)' with two PDF files listed: 'UBILLCable_081_1_1.1.pdf' and 'UBILLCable_082_1_1.1.pdf'. There are buttons for 'Update categories', 'Back', and 'Finish' at the top right.

_8. Select ASSIGN TO DOCUMENT TYPE

This screenshot is identical to the one above, showing the 'Create document types' page. The context menu over 'Category 1' now has 'Assign to document...' selected. The rest of the interface and document list are the same.

_9. Select Existing Document type then the appropriate **document type** from the drop-down list.



_10. Click Assign to close the dialog box

You can Click on any document to see a preview of it. This will help ensure the documents are correctly grouped.

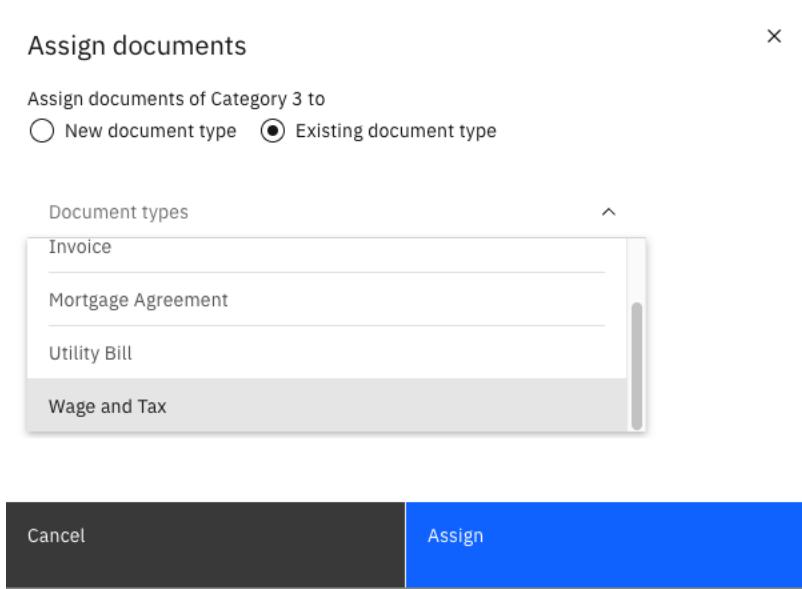
_11. Select the next Category 2 and Click on the 3 dots and Select Assign these documents to a document class.

_12. This time Select a New Document Type. Since we have not defined a mortgage agreement document type yet.

_13. Enter Mortgage Agreement in the field

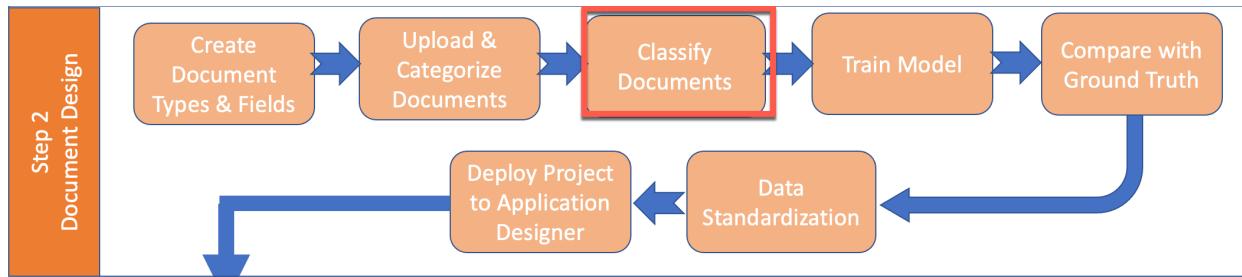
The screenshot shows the 'Assign documents' dialog box. At the top, it says 'Assign documents of Category 2 to'. Below that is a radio button for 'New document type' (checked) and another for 'Existing document type' (unchecked). The 'Document type display name' field contains 'Mortgage Agreement' (with character count 18/50). The 'Document type symbolic name' field contains 'MortgageAgreement' (with character count 17/50). At the bottom are 'Cancel' and 'Assign' buttons.

- _14. **Click Assign** to have the system automatically rename and move the category into the Document Types section.
- _15. Now for Category 3, **Click on 3 dots** and Select Assign Document type.
- _16. Select Existing Document Type and Click Wage and Tax from the drop down and then Click on Assign.



- _17. Once you confirmed all documents are correctly classified into the correct document type, **Click Finish**

7 Train classification



Now that we have documents uploaded in the system, we are ready to train the classification. Note that although you don't need a ton of document samples to train (minimum of 5), you are going to get better accuracy if the system has a deeper understanding of the documents, so more could be better.

In this lab, we curated some documents samples for you. In normal circumstances, you would need to do this yourself. Make sure the documents you upload to train classification are good documents.

- Clean documents
- High resolution
- Representative of the document type(s)
- Accurately grouped and uploaded to Document Processing

This is NOT the time to try and trick the system. Uploading a document that doesn't recognize well would not help the system recognize the types of words, phrases, and concepts it needs to learn to classify documents correctly.

The most common error is introducing a sample document into the incorrect document type, usually by uploading them to the wrong document type. If that happens, you are introducing conflict into the classification. For example, an invoice added to Tax Forms may confuse the system and result in it thinking invoices are tax forms and vice versa. Once that happens, you need to clean your documents and retrain the system.

- _1. Click on **<your project name>** in the cookie trail to return to the start page. In the example below our project was called **<User01_CEB>**
- _2. Click anywhere in the **CLASSIFICATION MODEL** line

Section	Status	Value	Details
Document types and samples	Ready	5 types	20 samples on average
Classification model	Ready	3 types trained	100% accuracy
Extraction model	Retrain	3 types trained	97% accuracy
Data standardization	Not ready		
Document retention	Ready	5 types reviewed	

Once we open the classification model, we will be presented with details on how to perform the retraining. There are four basic steps – Confirm inputs, Review Samples, Review Training Results, and Test Trained model.

On the Confirm inputs screen here we can confirm all the documents that will be used in this training exercise. We can also use the opportunity to remove documents that are no longer relevant or upload additional documents.

- _3. **Click Next** this will move from the Confirm inputs to the **Review Samples** step. Notice three documents have green icons next to them. These green icons show these documents have test samples already assigned. The new document types (Mortgage Agreement and Wage and Tax) do not have any test samples assigned yet therefore there's no green icons since we haven't assigned test sets yet.

Classification model

Accuracy 84.8%

Document types

- Bill of Lading
- Invoice
- Utility Bill
- Mortgage Agreement
- Wage and Tax

Mortgage Agreement sample documents (5)

Training/test ratio in % 100/0

Training set (5) 100% of total samples

Test set (0) 0% of total samples

There are no documents in the test set.

_4. For the Mortgage Agreement move two documents to the Test set by **checking** and **clicking on the arrow**.

Classification model

Accuracy 84.8%

Document types

- Bill of Lading
- Invoice
- Utility Bill
- Mortgage Agreement
- Wage and Tax

Mortgage Agreement sample documents (5)

Training/test ratio in % 60/40

Training set (3) 60% of total samples

Test set (2) 40% of total samples

_5. Select **Wage and Tax** on the Document types and move 2 documents over to the test set.

The suggested split is 70/30 – that is, 70% of the available sample documents should be used for training, and we will validate the training results with 30% of the sample documents. This split is only a suggestion, and we can adjust it, but 70/30 is a good starting point.

6. Select TRAIN to launch the training. This may take a several minutes. You will see a progress bar has training progresses.

Once complete, you will be able to see the training results.

What's happening: The samples are run through multiple machine learning algorithms. These machine learning algorithms learn from the ground truth, the association between the sample documents (the OCR text) and the document types. The yielding models are then evaluated with the documents in test set. The model-predicted document types on these documents are compared with the human-provided answers to compute the accuracy. The top three accurate models are presented to the user, with the most accurate one being selected by default.

You should see something like the following:

The screenshot shows the IBM Cloud Pak for Automation interface. At the top, it says "IBM Cloud Pak | Automation". Below that, it shows "Business automations / User01-CEB / Classification model". It says "Last trained: 4 minutes ago" and has an "Accuracy" badge of "96.9%" which is highlighted with a red box. There are four buttons at the top: "Confirm inputs", "Review samples", "Review training results", and "Test trained model" (optional). A message box on the right says "Model trained successfully!" and "Accuracy has been updated to reflect the latest changes." Below the buttons, there's a note: "Changes were made since you last trained your model. Retrain the model to get updated training results and accuracy." On the left, there's a sidebar with "Document types" and "Bill of Lading" selected. In the center, there's a "Training results" section with a table:

Document	Classified as	Classification result	Confidence
BOL_007.pdf	Bill of Lading	Correct	High
BOL_009.pdf	Bill of Lading	Correct	Medium
BOL_019.pdf	Bill of Lading	Correct	High
BOL_027.pdf	Bill of Lading	Correct	High
BOL_031.pdf	Bill of Lading	Correct	High
BOL_075.pdf	Bill of Lading	Correct	High

_7. Click on each of the document types. Notice the confidence levels. The both the Mortgage Agreement and Wage and Tax have a confidence of low (this will be pointed out even later after we deploy).

You can easily see where the system may be struggling. You should look for document types that don't match the actual file or have a low confidence. Remember the more documents you give to train, the better the results.

_8. Click on Next. This is the Test trained model. Here you can try and test other documents to see if they classified correctly. This step is optional but would be useful to try out the AI model to determine whether additional samples are necessary.

_9. Click Done

7.1 How do I improve my results?

Option 1 – Add more samples.

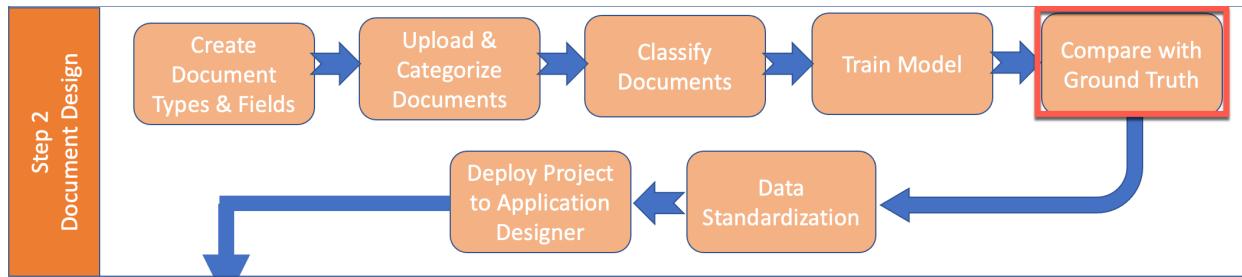
Option 2 – review all uploaded samples.

- remove those that are not a clear representation.
- remove those that are poor quality documents.
- carefully confirm that none of the samples contain multiple document types in the file. This is a common occurrence. A document is listed as a Purchase Order, but in the back pages, also contains other document types in that same file. This confuses the system.

To improve results, you would normally want to add more samples of the document ensuring they are clean and representative document to improve the system's understanding of the document.

Optional step add some more Wage and Tax documents from the Extra folder.

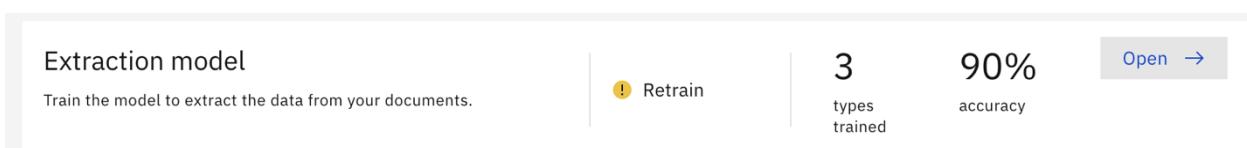
8 Data extraction



At this point, we have defined a document type, told the system which fields we want off the document and trained the system on how to recognize (classify) the document. In the Data Extraction portion of the lab, we will upload new Wage and Tax documents to Document Processing and see how our earlier configuration of the document type and related fields are working. This is comparing a new document extracted elements with the ground truth. Once we open Extraction model, we will be presented with details on how to perform the retraining. There are five basic steps – Review samples, Add fields, Teach the model, Review the trained model, and Test the model.

- _1. From the guided configuration screen, **Click** anywhere in the **Extraction model** box.

Note: the status will reset to Retrain if it detects something may have changed. This is just a reminder that if you indeed changed something, you may benefit from retraining the model.



- _2. Next **Click** on the **Wage and Tax** document type under the Document Types section.

Like in the classification step, ADP needs to have the documents divided into a training and test sets. In general, *deep learning*-based AI requires a larger number of sample documents to achieve a reasonable result. But since our environment does not have GPU deep learning is not turned on.

You should have something that looks like what you see in the following screen shot.

_3. Click on the **NEXT** button at the top.



You will now be on the Add fields bread crumb. If there were more fields to add we could do it here. But since we have already added all the fields needed, proceed to the next step.

_4. Click the **Next** button. You are now at the “Teach model” bread crumb.

Teach the model is where you will spend most of your time. We can see that our documents are “not ready” so we’ll need to teach the model with new documents.

_5. Click on **Teach Samples**.

Document name	Status	Fields reviewed	Date added
TR_FW2_1001_0000_PS.pdf	Not ready	0/7	09/29/2022, 11:03 am
TR_FW2_2000_0000_PS.pdf	Not ready	0/7	09/29/2022, 11:03 am
TR_FW2_3001_0000_PS.pdf	Not ready	0/7	09/29/2022, 11:05 am

Note: Your individual results may vary based on the exact documents you upload, how you configure your fields etc. Therefore, general guidance is given here versus exact step by step instructions.

_6. We will now review the fields that were extracted, correct any that may be wrong and add others.

You should now see the field data extracted by the system. Nothing has been trained yet. All it is doing is using the aliases we entered when we created the document class to locate data. Now, you need to correct and improve the model.

The screenshot shows the IBM Cloud Pak Administration interface with the following details:

- Document View:** The left pane displays the W-2 Wage and Tax Statement form for 2020. The form includes fields for Employee Social Security Number (123-45-6789), Employer Identification Number (14-023285), Employee Name and Address (Michael Robert David Smithson III, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234), and various tax withholdings and wages.
- Extracted Fields:** The right pane lists the extracted fields with their captured values. The first few entries are:
 - Federal Income Tax Withheld: 123456789.99
 - Social Security Wages: 123456789.99
 - Medicare Wages and Tips: 123456789.99
 - Social Security Tax Withheld: 123456789.99
 - Allocated Tips: 123456789.99
 - Dependent Care Benefits: 123456789.99
 - Nonqualified Plans: 123456789.99
 - State Income Tax Withheld: 123456789.99
 - Local Income Tax Withheld: 123456789.99
 - Loyalty Plan: ABCDEFGH
- Actions:** Buttons for "Save selection" and "Next sample" are visible at the bottom of the right pane.

Note: You may see different results than shown on the image above.

Let's spend some time showing how to go about correcting these issues to help the system learn how to extract the values accurately.

8.1 Correcting extracted values

Let's start with the Federal Income Tax withheld field (i.e., The first one in the 'Fields to extract' list). You will see that there are a series of blue underlines below all the characters found. We are interested in getting the "Federal Income tax withheld" data.

_1. Click on the number below the heading “Federal Income tax withheld” in the image.

The screenshot shows the IBM Cloud Pak Administration interface with the W-2 Wage and Tax Statement form loaded. A pop-up window titled "Save match" is displayed, prompting the user to save the captured value "123456789.99" under the field "Federal Income Tax Withheld". The "Save match" button is highlighted in blue.

_2. A pop-up window will ask if you want to save match of value captured along with the field label. **Select Save match**

Notice a green check mark signifies this field is complete.

The screenshot shows the IBM Cloud Pak Administration interface with the W-2 Wage and Tax Statement form loaded. The "Federal Income Tax Withheld" field now has a green checkmark next to its value "123456789.99", indicating it is saved. The "Save selection" button is visible at the bottom right of the sidebar.

The 3 ellipses next the green check mark allow you to clear the data or update ADP to there is no field with this data in the current view.

_3. Moving to Employee Name and Address field. Notice there are no blue lines under the actual name but there are blue lines for the Field label ("Employee's first name and initial"). Again, Click on Save match

The screenshot shows the IBM Cloud Pak Administration interface with the following details:

- Document:** TR_FW2_1001_0000_PS.pdf
- Fields Extracted:**
 - Federal Income Tax Withheld:** Value Captured: abc 123456789.99
 - Employee Name and Address:**
 - Field label (optional):** e Employee's first name and initial
 - Field value:** Draw
- Form Fields:**
 - Employer's social security number:** 577-22-3048
 - Employer identification number (EIN):** 14-023285
 - Employer's name, address, and ZIP code:** Long Lengthy Name The Corporation, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234
 - Control number:** 123456 A78
 - Employee's name and initial:** Michael Robert David S. (Captured field)
 - Employee's address and ZIP code:** 56334 Full Sized Avenue, Minneapolis, Minnesota
 - State:** MN
 - State tax ID number:** 123456789
 - State wages, tips, etc.:** 123456789.99
 - State income tax:** 123456789.99
 - Local wages, tips, etc.:** 123456789.99
 - Local income tax:** 123456789.99
 - Locality name:** ABCDEFGHI
 - Nonqualified plans:** 123456789.99
 - Retirement savings plan:** 123456789.99
 - Medicare wage and tips:** 123456789.99
 - Medicare tax withheld:** 123456789.99
 - Social security wage:** 123456789.99
 - Social security tax withheld:** 123456789.99
 - Social security tips:** 123456789.99
 - Allocated tips:** 123456789.99
 - Dependent care benefits:** 123456789.99
 - Other:** AAA BBB CCC 12345678.90, AAA BBB CCC 12345678.90
 - Subfields:** 16 items, 0 required items
 - Pending aliases:** None
- Buttons:** Save selection

The field label has been populated but we still need the field value.

_4. For the field value **Click** on the Draw button under Field value. Using your mouse **select** the Name and address (green box), then **Select Save selection**

The screenshot shows the IBM Cloud Pak Administration interface. On the left, a PDF of a W-2 form is displayed. The form includes fields for Employee Social Security Number (577-22-3048), Employer Identification Number (EIN) (14-023285), Employee Name and Address (Long Lengthy Name The Corporation, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234), Control Number (123456 A78), and various tax-related fields. A red box highlights the employee name and address. On the right, the extracted fields are listed in a table:

Field Name	Value Captured
Federal Income Tax Withheld	abc 123456789.99 Required
Employee Name and Address	Michael Robert David Smithson III 56334 Full Sized Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
Field label (optional)	Draw <input type="button" value="X"/>
Field value	Draw <input type="button" value="X"/>
Pending aliases	View all aliases (5)
None	<input type="button" value="None"/>
<input type="button" value="Save selection"/>	
Employee Social Security Number	abc Text
Employer Identification Number	abc Text
<input type="checkbox"/> Mark this document as ready for training.	<input type="button" value="Mark this document as ready for training."/>
<input type="button" value="Previous sample"/>	<input type="button" value="Next sample"/>

- _5. For the Employee Social Security field **Click on the number** then **Select Save selection**.
- _6. Continue to process for the remaining fields, using either method as described above, clicking on the blue lines or drawing a box around needed value.
- _7. Once complete **check the box** next to “Mark this document as ready for training” at the bottom

The screenshot shows the IBM Cloud Pak Administration interface. On the left, a PDF of a W-2 form is displayed. The form includes fields such as Employer Identification Number (EIN), Employee Name and Address, Social Security Number, and various wage and tax amounts. On the right, a table lists the extracted fields with their corresponding captured values. A red arrow points from the 'Field Value' section of the table to the 'Mark this document as ready for training' checkbox at the bottom of the interface.

Field Name	Value Captured
Federal Income Tax Withheld	abc 123456789.99
Employee Name and Address Required	
Employee Social Security Number Required	abc 577-22-3048
Employer Identification Number	abc 14-023285
Employers Name and Address	abc Long Lengthy Name The Corporation 56334 Full ...
Social Security Wages	abc 123456789.99
Wages Tips Other Compensation	abc 123456789.99

Mark this document as ready for training. [\(i\)](#)

- _8. Review ALL other fields carefully. **Do not leave any incorrect values.** You can adjust or delete values as needed by clicking on Edit selection. If you leave incorrect values, the system will assume they are correct and actually LEARN them as if they were good values.
- _9. Repeat **steps for Next Sample**
Over the course of next few samples you may find that ADP has extracted the wrong results, perhaps getting a value that is above when it should have been below. If this is the case and you pick you a blue underline data, but the results are wrong. Simply use the draw box for the Field Label and Field Value.
- _10. Once complete review of all the sample documents **Click** on the **Back link**

The screenshot shows the IBM Cloud Pak Administration interface. On the left, there are two W-2 form documents. The top one is titled "2020 W-2 and EARNINGS SUMMARY". It contains various fields such as Employee Name (Michael Robert David Smithson III), Social Security Number (123456789.99), and wages (123456789.99). The bottom W-2 form is identical. On the right, there is a "Field Name" table with columns "Field Name" and "Value Captured". The table includes entries like "Federal Income Tax Withheld" (123456789.99), "Employee Name and Address" (Long Lengthy Name The Corporation), and "Social Security Wages" (123456789.99). Below this table is a "Wages Tips Other Compensation" section with a "Field value" input field containing "123456789.99". At the bottom, there is a checkbox "Mark this document as ready for training." and a "Save selection" button.

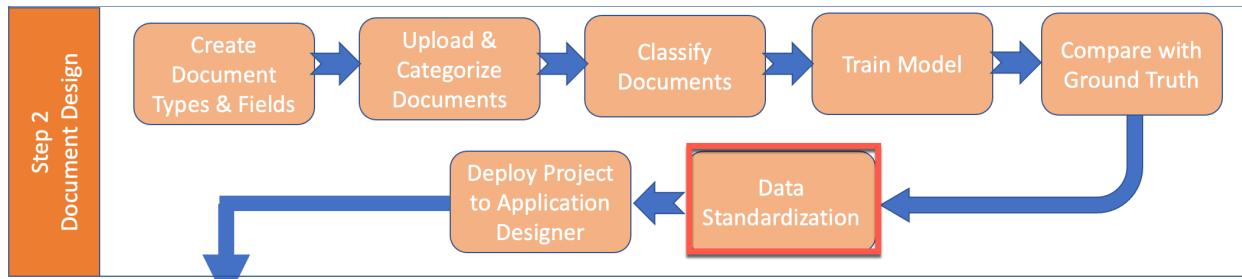
8.2 Train extraction model

We will be performing the quick training in this lab due not having a GPU in our TechZone architecture. A GPU is only needed a development environment and is not needed in either a production or runtime environment. The Deep Learning capabilities have been disabled on this training environment. You can find instructions in the Appendix for when you have access to a server with it enabled.

1. Click Train button.

This will take several minutes. (Good time for a break)

9 Data standardization

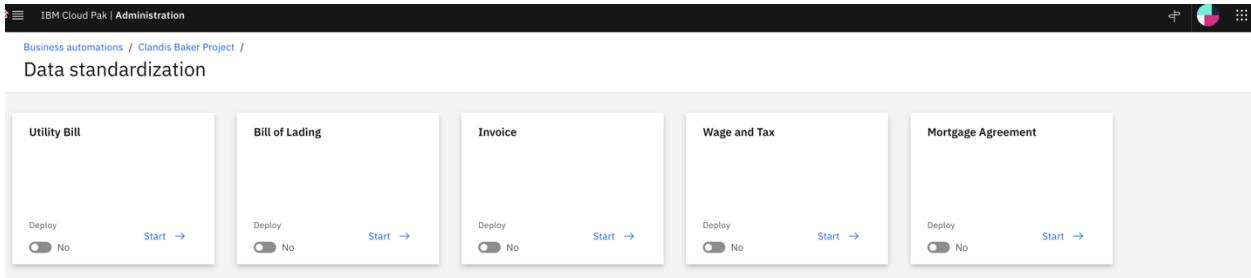


Next, we need to standardize the data that will be presented in the user interface and how it will be stored in the FileNet repository for example. Data standardization is the process of defining attributes for a data field in a standardized way. This is done using data definitions. These definitions can be used across projects, and across different applications within the CloudPak for Automation. Each data definition has a title, description, and a datatype. We can also set a data definition as required or not. When a document is ingested into ADP, it results in a list of Key Value Pairs' (KVP) for that document. The Designer maps some of these KVP's to fields and teaches the model on how to extract the fields from the full list of KVP's. The designer then maps some of those fields to data definitions for a particular document type. Only the fields that have been mapped to data definitions will become Content Process Engine properties.

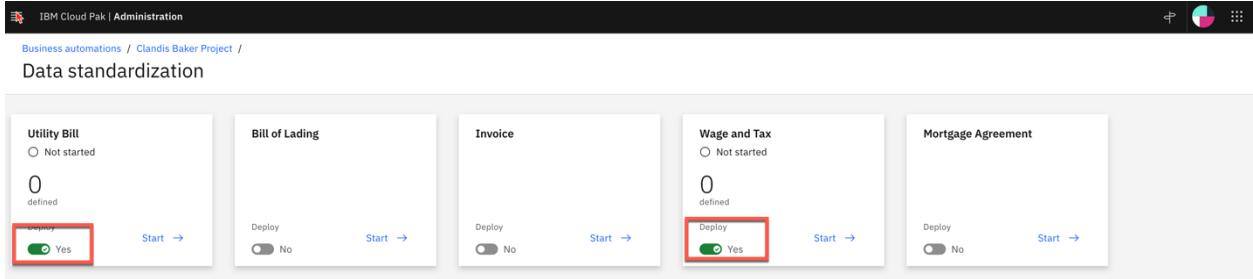
- _1. Return to the guided configuration flow and **Click** anywhere in the **Data standardization** box

Document types and samples	<input checked="" type="radio"/> Ready	4 types	22 samples on average
Upload sample documents to define the types of documents you want the system to process.	→		
Classification model	<input checked="" type="radio"/> Ready	3 types trained	100% accuracy
Train the model to classify your documents.	→		
Extraction model	<input checked="" type="radio"/> Ready	3 types trained	97% accuracy
Train the model to extract the data from your documents.	→		
Data standardization	<input type="radio"/> Not ready		
Map fields to new or existing data definitions.	→		

Here, you will see a list of available document types. Only the ones which have Deployed turned on will be visible in the verify interface and will have fields stored in FileNet.



_2. Ensure the Utility Bill and Wages and Tips and Deploy is toggled to Yes



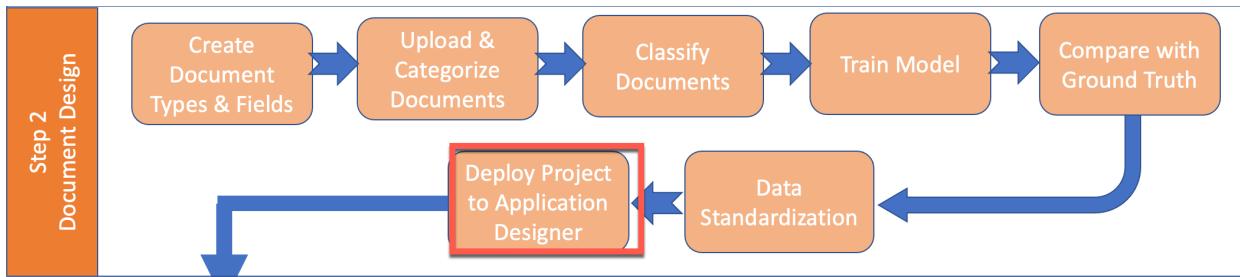
_3. Click on Start on either selected deployment.

This is where we begin defining the data file attribute definitions. You could create a new data definition and configure them.

_4. Return to the guided configuration screen by Clicking on <your project> name at the top of the screen.

Business automations / Clandis Baker Project /

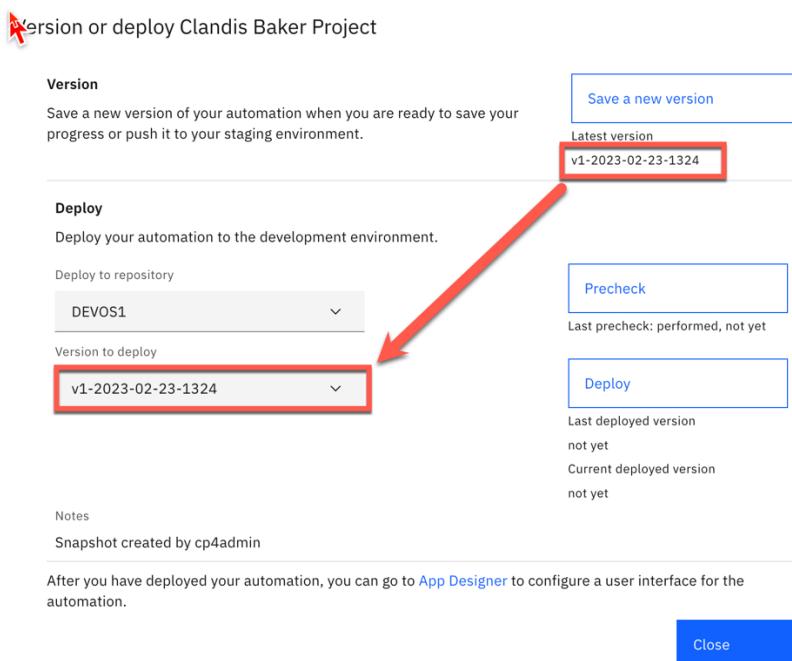
10 Version and deploy your project



At this point in our Designer project, we have defined a document type, labeled the fields we want from the document, trained (classified) the system to recognize the document type, reviewed the extracted fields we wanted and standardized (mapped) the document fields to our output.

Now that we completed the configuration of the content extraction project, we need to save and deploy the design project to the application side. This will allow you to test your project using a client runtime interface.

- _1. If not already there, return to the guided home screen by clicking on your project name. Then **Click Version / Deploy**
- _2. Click **Save a new version**.
- _3. Once the version is saved, you should see the version in the Version to deploy drop down list



... also, in the top corner has the “Latest Version”

4. Click on the **Deploy button**. This will also take several minutes and potentially time out if others are also trying to deploy.

Once completed, you should have a notice that the project was deployed.

The screenshot shows the 'Version or deploy Clandis Baker Project' dialog. In the 'Deploy' section, the 'Deploy to repository' dropdown is set to 'DEVOS1'. The 'Version to deploy' dropdown is set to 'v1-2023-02-23-1324'. The 'Deploy' button is highlighted with a red box. Below it, a message box displays the last deployed version as 'v1-2023-02-23-1324' and the current deployed version as 'v1-2023-02-23-1324'.

Note that you do not have to remain in the deploy screen while it is versioning or deploying. You can always click the button and then go back into any other screen if you like. It will run in the background. If you do this, just keep an eye on the top right of your screen for deployment status.

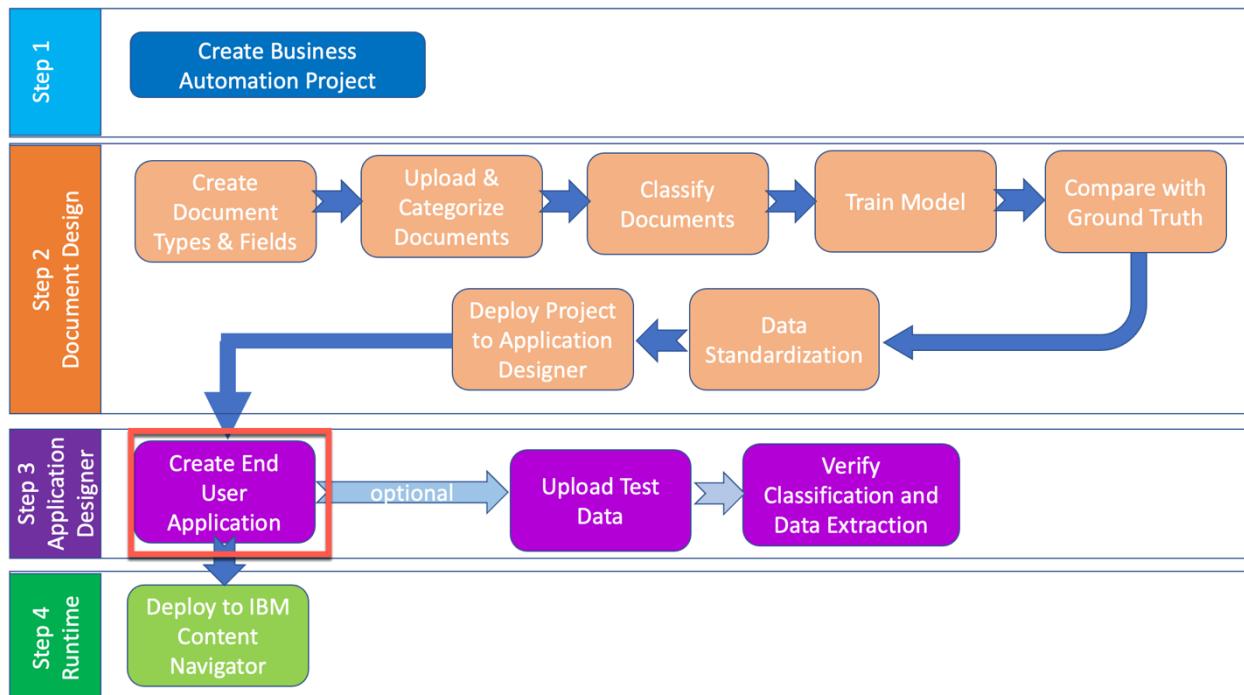
5. Click **Close** button.

Once deployed, proceed to the next steps.

From the home screen you can see the latest version and deployment

The screenshot shows the 'IBM Cloud Pak | Administration' home screen for the 'Clandis Baker Project'. In the top right corner, there is a 'Version / Deploy' button. To its right, a message box displays 'Latest version | v1 | 13 minutes ago' and 'Deployed | v1 | 6 minutes ago', both highlighted with a red box.

11 Application designer



At this point we have designed or built a project that consists of document types, data or file types and methods to extract the desired data. The next major section of this lab is to build the user interface using the Application Designer. IBM provides two application templates for Document Processing

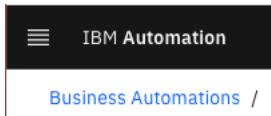
1. Batch Document Processing template – used to process batches of documents.
2. Document Processing Template – used to process single documents.

The lab will have you create a new batch processing application. We will quickly explore the various tabs in the interface, preview what the IBM Content Navigator (ICN) client would look like using the Preview feature and then publish our application to ICN where we will process a batch of documents.

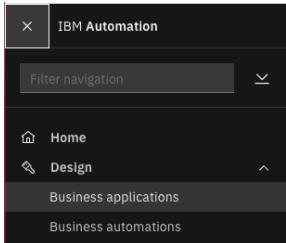
Changes to the application itself will not be in the scope of this lab.

11.1 Create your Runtime Application.

- _1. Return to the starting screen by **clicking the hamburger** in the top left.



and **selecting Business Applications**



_2. From the **Create** drop down list, select Application

Business applications

Quickly create user interfaces that integrate tasks, data, and automations. You can start with a template to ensure consistency. You can also use toolkits to share artifacts from existing applications. [Learn more](#)

Request Approval template	Onboarding Application template	Exception Handling template
Use this template to create a service desk request.	Use this template to onboard new employees to your organization.	Use this template to create a basic refund request application.
Last updated 02/20/2023	Last updated 02/20/2023	Last updated 02/20/2023

_3. Select Enter your <application name> in the Name field.

Create a business application

Name

Clandis Baker Application

Purpose (optional)

Describe the purpose of the application

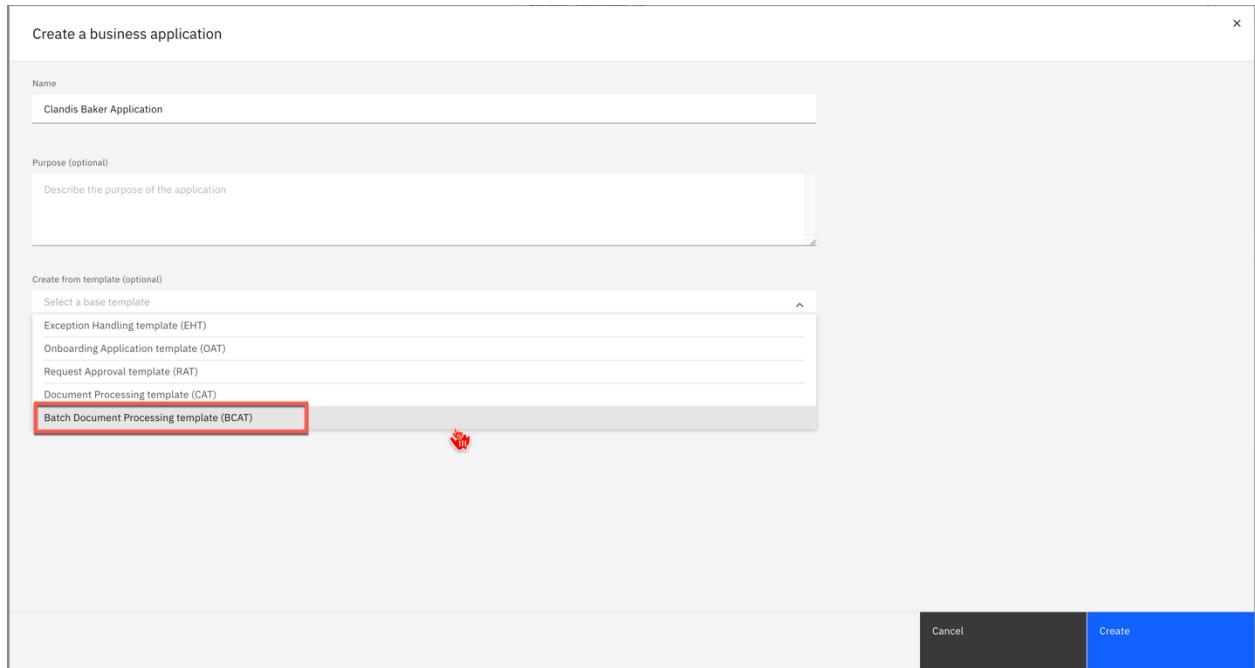
Create from template (optional)

Select a base template

- Exception Handling template (EHT)
- Onboarding Application template (OAT)
- Request Approval template (RAT)
- Document Processing template (CAT)
- Batch Document Processing template (BCAT)

Cancel Create

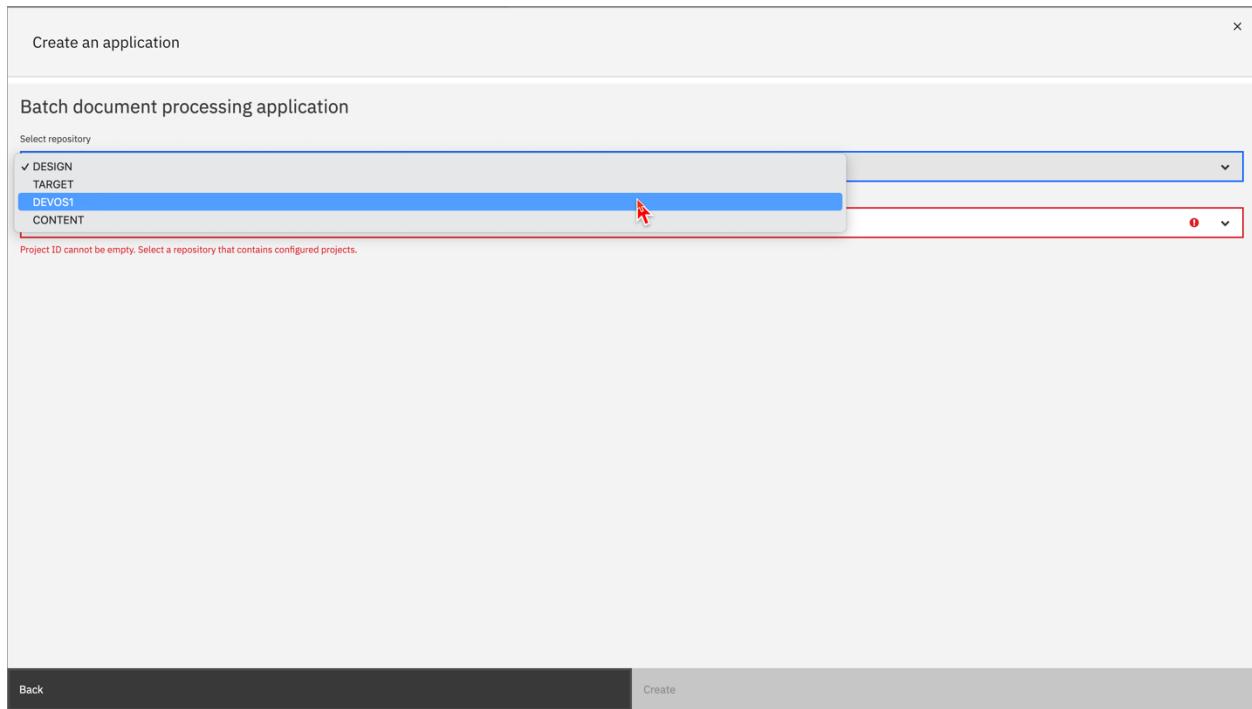
- _4. In the Create Form Template give it a <Name> and in drop down **select Batch Document Processing template (BCAT)**.



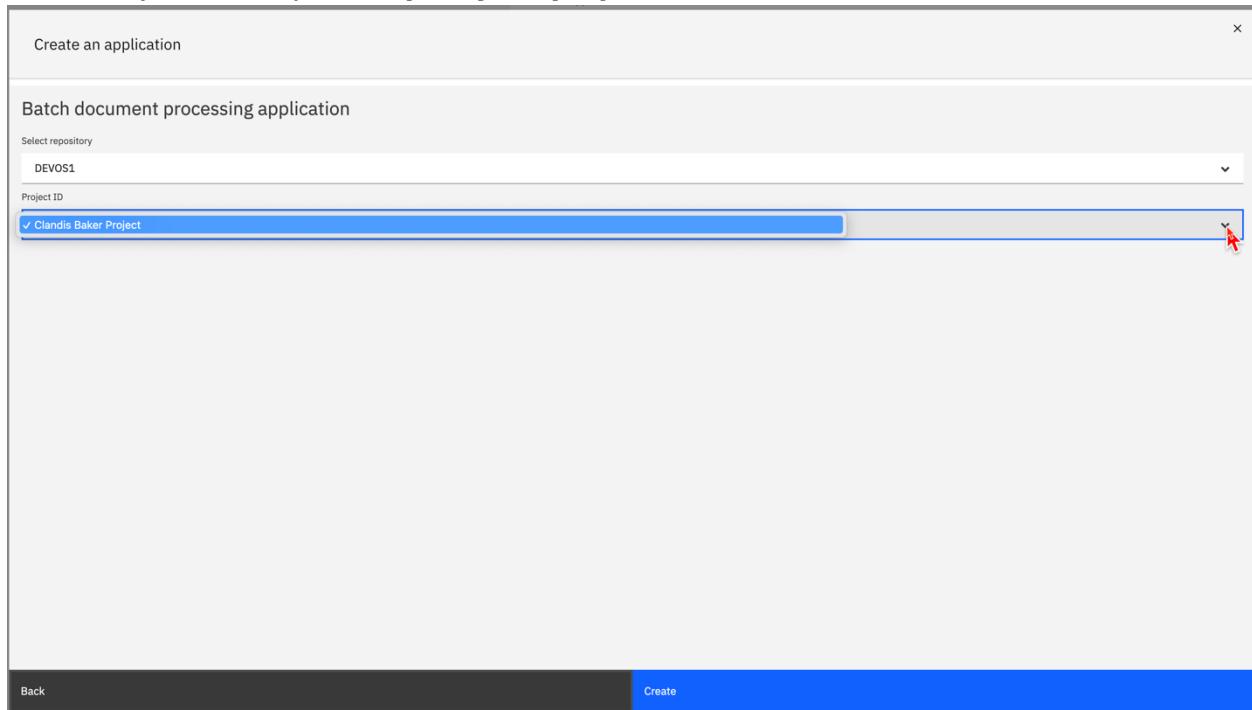
You could have selected the Document Processing Template if you only wanted to process a single document at a time, but in this lab, you will process several documents in a batch.

- _5. Click **Next**

- _6. You will be presented with the Create an application window. In the Select repository **pick DEVOS1**



_7. In the Project ID drop down **pick your project name**.



_8. Click **Create**

You should now be in the *Application Designer*

The screenshot shows the IBM Cloud Pak Application Designer interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Administration' and a 'Preview' button. Below the navigation, the title 'Clandis Baker Application' is displayed. The main content area shows 'Review batch issues' with two sections: 'Document type and page order issues' and 'Data extraction issues'. Below this, a 'Content List' section titled 'Batches' shows a table with three rows of document data:

Name	Size	Modified by	Last modified	Version
My Document1	2 KB	User1	10/1/2022, 01:10 AM	1
My Document2	1 MB	User2	10/2/2022, 02:20 AM	2
My Document3	90 B	User3	10/3/2022, 03:30 AM	3

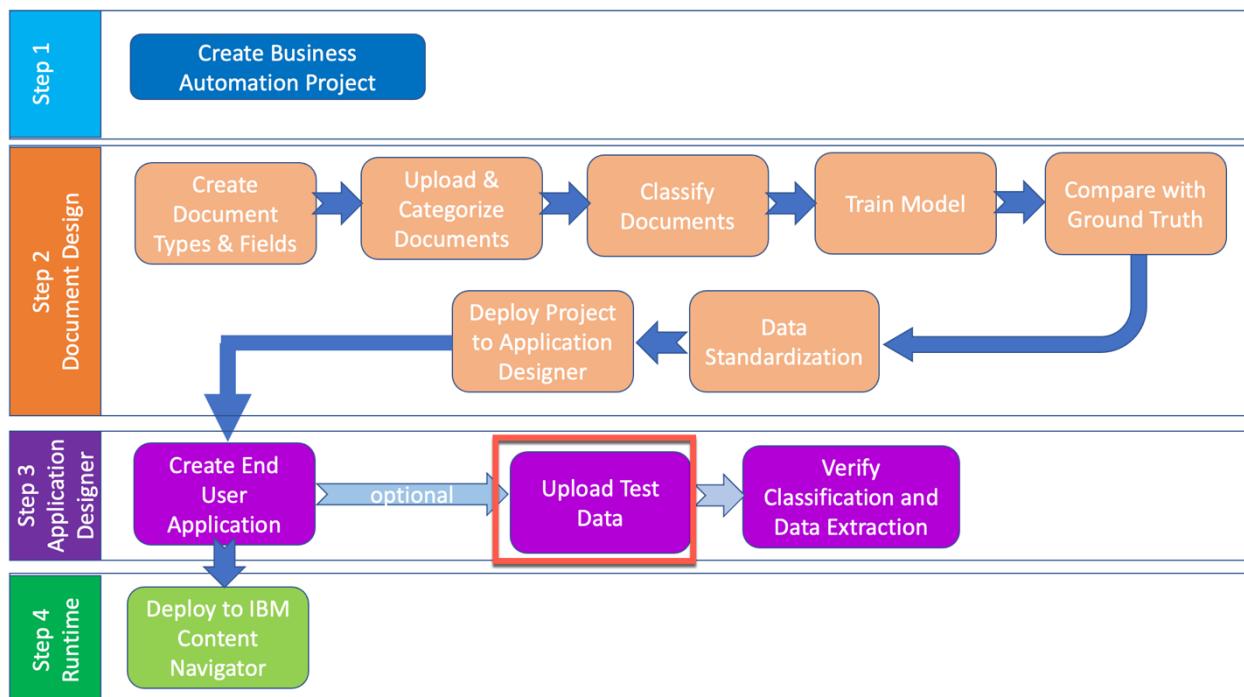
At the bottom of the content area, there are buttons for 'Items per page: 100' and 'Items 1-3'. To the right of the content area is a sidebar titled 'Drag a component to your page' containing a grid of UI component icons. The components include: Add batch modal, Add document modal, Add folder modal, Batch content, Button, Check box, Collap... panel, Content list, Content properties, Data verify..., Date/time picker, Decimal, Delete object modal, Display text, Document correction, Document reference, Document thumbnail, Document viewer, Edit prop... entries m..., Export document.

Batch Document Processing template (BCAT) has all the necessary pages and configuration to start using the application. Using this designer user interface, you have the option to further customize the application, such as its page design or actions, to fit your requirements.

_9. Click **Preview** at the top right corner.

Note: It may take several seconds to build and display the current configuration of the interface.

11.2 Upload documents for processing

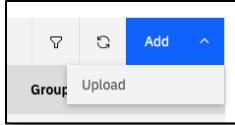


_1. You should be in the default application user interface for ADP.

Name	Files	Priority	Status	Added on	Added by	Group	Location
<input type="text"/> Add No items found.							

There are two key screens you will work with: “*Document type and page order issues*” and the “*Data extraction issues*”. First, we need to upload some test documents and have them processed.

_2. Click on Add, then Upload.



- _3. Enter a **name** for your batch in the Display Name field and set the **Priority to High** as seen in the image below.

Upload new batch

* Display Name
Batch 1

Description

Priority
High

- _4. Click **Select files**.

Navigate to the samples folder previously downloaded and use the *Group 2 ADP Application* folder documents.

- _5. Click **Open**

You will see a window that will give the operator a chance to manually classify the documents before they are ingested. In the example below would be how to manually classify a document. We are not going to do this but instead let ADP auto classify them.

Add Files

To manually specify document type, first select the files in the table. Use the classify option, to assign the document type for selected file(s). If a file is not manually classified, the system will auto-classify it.

1 items selected		Classify ^	Auto Classify	Deselect
<input type="checkbox"/>	File Name	Utility Bill		
<input checked="" type="checkbox"/>	TR_FW2_1001_0001_PS.pdf	Wage and Tax		
<input type="checkbox"/>	TR_FW2_1001_0002_PS.pdf	Auto Classify		
<input type="checkbox"/>	TR_FW2_2000_0001_PS.pdf	Auto Classify		
<input type="checkbox"/>	TR_FW2_3001_0001_PS.pdf	Auto Classify		
<input type="checkbox"/>	TR_FW2_4000_0009_PS.pdf	Auto Classify		

Cancel Add

_6. Click on the Add button.

Review batch issues

Batches

Name	Files	Priority	Status	Added on	Added by	Group	Location
Batch01	5	High	<div style="width: 60%;">3 of 5 files processed</div>	02/23/2023, 10:49 AM	cp4admin		

Items per page: 100 1-1 of 1 items

A progress bar will be displayed indicating when all documents have been uploaded.

_7. Click the 3 dots at the end of the line.

Review batch issues

Batches

Name	Files	Priority	Status	Added on	Added by	Group	Location
Batch01	5	High	Documents uploaded	02/23/2023, 10:49 AM	cp4admin		

Items per page: 100 1-1 of 1 items

_8. Click Submit

In the screen shot below, you see we have a document issues (status) and we now have 1 batch in the “Document type and page order issue” tile.

Review batch issues

Batches

Name	Files	Priority	Status	Added on	Added by
Batch 1	6	High	Document issues	01/13/2021, 08:44 am	CEAdmin

Items per page: 100 1-1 of 1 items

11.3 Correct any classification errors.**_1. Click on the Document type and page order issues tile to open the batch.**

Batch Document Processing Application /

Document type and page order issues

Name	Priority	Status	Added on	Added by	Group	Location
Batch 1	High	Document issues	01/13/2021, 08:44 am	CEAdmin		

Items per page: 100 | 1-1 of 1 items

_2. Click on <your batch name> to open it.

You should now see all the documents you uploaded in your batch. The ones with issues will have a yellow checkmark for documents that have a low confidence document type and a red exclamation mark for documents it could not classify.

Batch01

Cancel | Save changes | Submit

Documents (5)	Add +
Issues (1 of 5)	Dismiss
Document name	Document type
Review document type TR_FW2_1001_0001_PS.pdf	Wage and Tax
Review document type TR_FW2_1001_0002_PS.pdf	Wage and Tax
Review document type TR_FW2_2000_0001_PS.pdf	Wage and Tax
Review document type TR_FW2_3001_0001_PS.pdf	Wage and Tax
Review document type TR_FW2_4000_0009_PS.pdf	Utility Bill

Edit actions for .PDF and .TIFF

The screenshot shows a list of five documents under the 'Batch01' batch. The first four documents are flagged with a yellow warning icon and labeled 'Review document type'. The fifth document is flagged with a blue error icon and labeled 'Utility Bill'. To the right, a preview window displays a 'W-2 Wage and Tax Statement' form for the year 2020. The form includes fields for employee information, wages, taxes, and deductions. The preview window has a zoomed-in view of the top left corner of the form.

Why did all of Wage and Tax get flagged for review of document type? If you remember back in the classification section, we only uploaded the bare minimum of 5 documents and our classification was marked low. By adding more documents, we can train ADP further and not receive low confidence on these documents.

_3. Most of the document types are correct so we can Click on Dismiss

_4. If the last document has the wrong Document Type. Click on the Pencil icon and Select Wage and Tax then Select Update

Batch01

Document name	Document type
TR_FW2_1001_0001_PS.pdf	Wage and Tax
TR_FW2_1001_0002_PS.pdf	Wage and Tax
TR_FW2_2000_0001_PS.pdf	Wage and Tax
TR_FW2_3001_0001_PS.pdf	Wage and Tax
TR_FW2_4000_0009_PS.pdf	Wage and Tax (Recommended)

Select document type

Document type: Wage and Tax (Recommended)

Cancel Update

Form W2 Wage and Tax Statement
Copy B - To Be Filed with Employee's FEDERAL Tax Return
OMB No. 1345-0008
Year 2020
Form W2 Wage and Tax Statement
Copy C - For Employer's Records
OMB No. 1345-0008
Year 2020

1

Page 1 / 1

- _5. Review all documents to ensure everything is correct. If the system no longer detects any issues, you should see a green checkmark near the top of the document list.



- _6. Click **Submit** to save your changes and have the batch processed.

The system will start reprocessing the documents now that they have been classified correctly.

- _7. Click on the **Batch Document Processing Application** link at the top to return to the previous preview menu.

Batch Document Processing Application /
Document type and page order issues

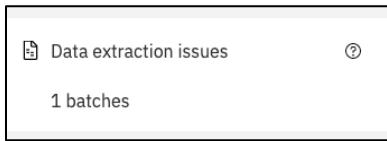
11.4 Correct extraction issues

The following instructions are based on a pre-trained sample application. Not what you will see in your untrained application.

Important Note: The project you are using for this has been configured but NOT run through the training (Deep Learning). So, the results will not reflect what they should be. IN A NORMAL SCENARIO, ON A CLUSTER WITH GPU AND DEEP LEARNING ENABLED, YOU WOULD HAVE TRAINED YOUR MODEL BEFORE DEPLOYING IT AND WOULD BENEFIT FROM HIGHER EXTRACTION RATES. the purpose of this lab is to teach you the tools but won't show you the trained results.

It may take a few seconds for your batch to advance to the next step. If your batch needs further attention, you will see it appear in the Data extraction issues tile.

_1. Click on the **Data extraction issues** tile to open it.



_2. Click on <your Batch name> to open.



After opening we see all the documents that have been processed but have extraction issues.

Batch Document Processing Application / Batches with data extraction issues /				
Name	Issues	Status	Modified on	Modified by
TR_FW2_1001_0001_PS.pdf	1	⚠ Data issues	23/02/2023	cp4admin
TR_FW2_1001_0002_PS.pdf	1	⚠ Data issues	23/02/2023	cp4admin
TR_FW2_2000_0001_PS.pdf	1	⚠ Data issues	23/02/2023	cp4admin
TR_FW2_3001_0001_PS.pdf	2	⚠ Data issues	23/02/2023	cp4admin
TR_FW2_4000_0009_PS.pdf	0	Issues reviewed	23/02/2023	cp4admin

Items per page: 100 1-5 of 5 items

Notice 4 of the 5 documents have Data issues. One document has 2 issues raised. And the last one doesn't have any. What happened? Why are we getting document issues on most of our documents? The reason again is our low confidence for the classification of Wage and Tax.

_3. Click on the first document to open it. Notice the yellow triangle at the top.

TR_FW2_1001_0001_PS.pdf | Document type: Wage and Tax | Classification: Low confidence.

Extracted data

All Fields

Federal Income Tax Withheld *

Federal Income Tax Withheld
1800.00

Similar fields

Employers Name and Address

Test and Rest Inc. 563 Stoney Brook Rd Minneapolis, MN 55411

Social Security Wages

17700.00

Wages Tips Other Compensation

18000.00

Employee Social Security Number *

577-22-3048

Employer Identification Number *

14-023285

Employee Name and Address *

Organization
(none)

Name
(none)

The screenshot shows a W-2 form for employee 22222 with SSN 577-22-3048. The form includes fields for wages, tips, other compensation (18000.00), federal income tax withheld (1800.00), social security wages (17700.00), social security tax withheld (1113.33), Medicare wages and tips (18000.00), Medicare tax withheld (261.00), state wages, tips, etc. (18000.00), state income tax (1260.00), local wages, tips, etc. (17700.00), local income tax (500.00), and MPLS. The employer information is Test and Rest Inc. at 563 Stoney Brook Rd, Minneapolis, MN 55411. The employer ID is 14-023285 and the employee SSN is 577-22-3048. The document is dated 2020.

Take a moment to discover the image viewer features.

Image viewer features at top:

TR_FW2_1001_0001_PS.pdf | Document type: Wage and Tax | Classification: Low confidence.

Extracted data

Fields with issues

Employee Social Security Number *
Validation error
577-22-3048

Employer Identification Number
Validation error
14-023285

The screenshot shows a W-2 form for employee 22222 with SSN 577-22-3048. The form includes fields for wages, tips, other compensation (18000.00), federal income tax withheld (1800.00), social security wages (17700.00), social security tax withheld (1113.33), Medicare wages and tips (18000.00), Medicare tax withheld (261.00), state wages, tips, etc. (18000.00), state income tax (1260.00), local wages, tips, etc. (17700.00), local income tax (500.00), and MPLS. The employer information is Test and Rest Inc. at 563 Stoney Brook Rd, Minneapolis, MN 55411. The employer ID is 14-023285 and the employee SSN is 577-22-3048. The document is dated 2020. There are validation errors for the Employee Social Security Number and Employer Identification Number.

- Rotate image
- Visual effect adjustment
- Invert

Image viewer features at bottom:

The screenshot shows a document processing interface. On the left is a large preview window displaying a W-2 form. On the right is a panel titled "Extracted data" showing various fields and their values. At the bottom is a toolbar with icons for zoom, fit-to-page, and search, along with a page navigation bar.

Extracted data

Fields with issues (2)

- Validation error Employee Social Security Number * 577-22-3048
- Validation error Employer Identification Number 14-023285

Employee's social security number	577-22-3048	OMB No. 1440-0008			
Employer identification number (EIN)	14-023285	1. Wages, tips, other compensation	18000.00	4. Federal income tax withheld	\$1000.00
Employer's name, address, and ZIP code	Test and Repair 563 Storey Brook Rd Minneapolis, MN 55411	2. Social security wages	17700.00	5. Social security tax withheld	1113.33
		3. Medicare wages and tips	18000.00	6. Medicare tax withheld	261.00
		7. Social security tips	400.00	8. Allocated tips	400.00
Control number	101220 A13	9. State nonqualified plans	300.00	10. Dependent care benefit plan	543.21
Employee's first name and initial	Benjamin P. Charles	11. Nonqualified plans	300.00	12. State income tax withheld	250.00
Last name	Aldrich	13. Federal income tax withheld	1260.00	14. Local income tax withheld	500.00
Address	4326 Aldrich Rd	15. State income tax withheld	20000.00	16. Local income tax withheld	MPLS
City, State, Zip code	Minneapolis, MN 55412	17. Local wages, tips, etc.	17700.00	18. City, state	
State	Minn.	19. State wages, tips, etc.	18000.00	20. Local income tax withheld	500.00
ZIP code	795037	21. Local wages, tips, etc.	1260.00	22. Local income tax withheld	MPLS

Form W-2 Wage and Tax Statement 2020 Department of the Treasury - Internal Revenue Service
Copy 1 --For State, City, or Local Tax Department

Cancel Save changes Done and next Done

- Page and thumbnail's view
- Fit to window
- Zoom and Magnify

Field features

The screenshot shows the same document processing interface as above, but with a red box highlighting the "Fields with issues" section in the "Extracted data" panel. This section lists validation errors for the Employee Social Security Number and Employer Identification Number.

Extracted data

Fields with issues (2)

- Validation error Employee Social Security Number * 577-22-3048
- Validation error Employer Identification Number 14-023285

Employee's social security number	577-22-3048	OMB No. 1440-0008			
Employer identification number (EIN)	14-023285	1. Wages, tips, other compensation	18000.00	4. Federal income tax withheld	\$1000.00
Employer's name, address, and ZIP code	Test and Repair 563 Storey Brook Rd Minneapolis, MN 55411	2. Social security wages	17700.00	5. Social security tax withheld	1113.33
		3. Medicare wages and tips	18000.00	6. Medicare tax withheld	261.00
		7. Social security tips	400.00	8. Allocated tips	400.00
Control number	101220 A13	9. State nonqualified plans	300.00	10. Dependent care benefit plan	543.21
Employee's first name and initial	Benjamin P. Charles	11. Nonqualified plans	300.00	12. State income tax withheld	250.00
Last name	Aldrich	13. Federal income tax withheld	1260.00	14. Local income tax withheld	500.00
Address	4326 Aldrich Rd	15. State income tax withheld	20000.00	16. Local income tax withheld	MPLS
City, State, Zip code	Minneapolis, MN 55412	17. Local wages, tips, etc.	17700.00	18. City, state	
State	Minn.	19. State wages, tips, etc.	18000.00	20. Local income tax withheld	500.00
ZIP code	795037	21. Local wages, tips, etc.	1260.00	22. Local income tax withheld	MPLS

Form W-2 Wage and Tax Statement 2020 Department of the Treasury - Internal Revenue Service
Copy 1 --For State, City, or Local Tax Department

Cancel Save changes Done and next Done

- Show all fields.
- Show fields with issues.

Also note that fields that do have issues have a notification icon next to them. For example, the Employee Social Security Number field is a mandatory field and expects a numeric value. But in this example this field also has hyphens in it therefore didn't pass validation.

The screenshot shows a document processing interface with a W-2 form on the left and an 'Extracted data' panel on the right.

W-2 Form Data:

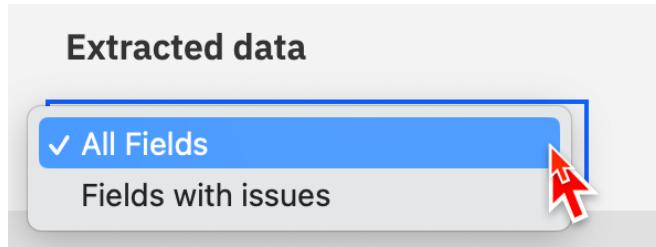
- Employee's social security number: 577-22-3048
- Employer identification number: 14-023285
- Employer's name, address, and ZIP code: Test and Rest Inc., 563 Storey Brook Rd, Minneapolis, MN 55412
- Control number: 210220 A13
- Employee's first name: Benjamin P. Charles
- Employee's address and ZIP code: 4326 Aldrich Rd, Minneapolis, MN 55412
- Form: W-2 Wage and Tax Statement, Copy 1—For State, City, or Local Tax Department, 2020

Extracted data Panel:

The 'Fields with issues' section highlights two validation errors:

- Employee Social Security Number ***: Validation error (577-22-3048)
- Employer Identification Number**: Validation error (14-023285)

_4. Under Extracted data click on the drop down twisty.



_5. Click on the **ALL Fields**.

This view shows all the fields that we defined earlier. Fields with an asterisk are mandatory fields.

TR_FW2_1001_0001_PS.pdf | ⚠ Document type: Wage and Tax | Cancel Save changes Done and next Done

Extracted data

All Fields

Federal Income Tax Withheld *
1800.00

Employers Name and Address
Test and Rest Inc. 563 Stoney Brook Rd Minneapolis, MN 55411

Social Security Wages
17700.00

Wages Tips Other Compensation
18000.00

Employee Social Security Number *
577-22-3048

Employer Identification Number *
14-023285

Employee Name and Address *

If we change the Extracted data back to Fields with issues:

TR_FW2_1001_0001_PS.pdf | ⚠ Document type: Wage and Tax | Cancel Save changes Done and next Done

Extracted data

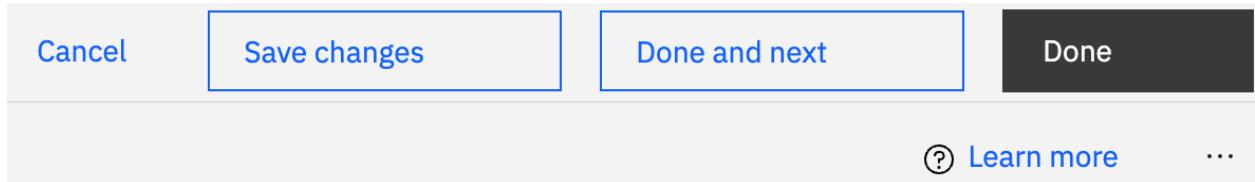
Fields with issues

Issue types

There aren't any extraction issues

Notice no fields are displayed since ADP was able to get all the mandatory fields required.

_6. Click on **Done and next** box at the top.



_7. For the next document there are no extraction issue only low confidence on document type. For this document you shouldn't have any issues to resolve.

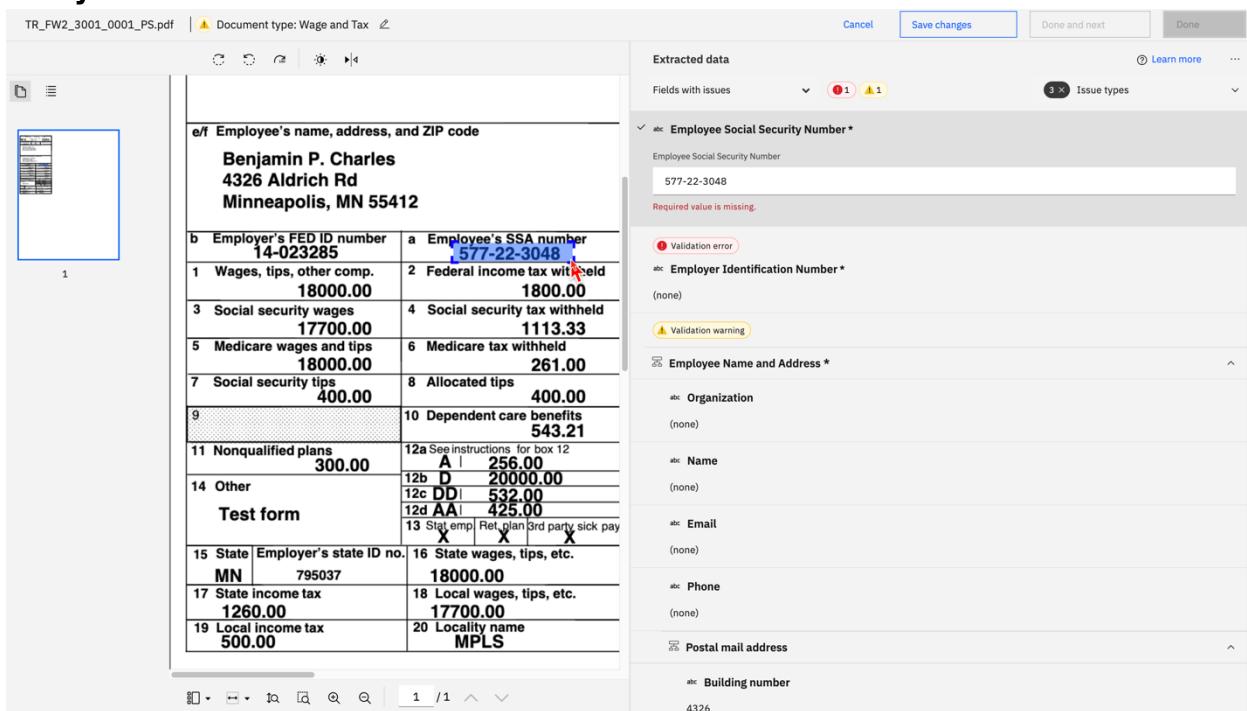
_8. Click on **Done and next** again. And again, no issues with our mandatory fields.

_9. Click on **Done and next** again. Now we are at the document which earlier in the queue told us there were 2 issues (step 2 above).

The screenshot shows a document viewer window with a W-2 form. To the right is an 'Extracted data' panel. The panel has sections for 'Employee Social Security Number', 'Employer Identification Number', and 'Employee Name and Address'. Each section has a validation error message and a validation warning message. The 'Fields with issues' dropdown shows 2 validation errors and 1 validation warning.

_10. Click on the Employee Social Security Number.

You may have to zoom in a bit so you can see where the SSA number is located.

_11. Take your mouse and lasso around the SSN number.


The screenshot shows a document processing application interface. On the left is a preview of the PDF form, which contains a table of wage and tax information. On the right is the 'Extracted data' panel. In the preview, the Employee Social Security Number field (line 1) is highlighted with a blue box. In the extracted data panel, the same field is also highlighted with a red border, indicating it has been selected or is being processed. The panel lists various fields and their extracted values, including the Employee Social Security Number (577-22-3048), which is marked as required and missing.

b	Employer's FED ID number 14-023285	a	Employee's SSA number 577-22-3048
1	Wages, tips, other comp. 18000.00	2	Federal income tax withheld 1800.00
3	Social security wages 17700.00	4	Social security tax withheld 1113.33
5	Medicare wages and tips 18000.00	6	Medicare tax withheld 261.00
7	Social security tips 400.00	8	Allocated tips 400.00
9		10	Dependent care benefits 543.21
11	Nonqualified plans 300.00	12a	See instructions for box 12 A 256.00
14	Other Test form	12b	D 20000.00
		12c	DD 532.00
		12d	AA 425.00
15	State Employer's state ID no. MN	16	State wages, tips, etc. 18000.00
17	State income tax 1260.00	18	Local wages, tips, etc. 17700.00
19	Local income tax 500.00	20	Locality name MPLS

Extracted data

Fields with issues (1) (1)

Employee Social Security Number
577-22-3048
Required value is missing.

Validation error
Employer Identification Number *
(none)

Validation warning
Employee Name and Address *

Organization
(none)

Name
(none)

Email
(none)

Phone
(none)

Postal mail address
Building number
4326

_12. Repeat same steps above for Employer Identification Number.**_13. Click Save Changes at the top.****_14. Select Done and next.****_15. All documents have been processed Click on Submit at the top to complete the batch.**

END OF LABS

12 Export Import Project.

From the Business Automations

1. From the Business Automations screen select Document Processing.

The screenshot shows the 'Business automations' section of the IBM Cloud Pak interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Administration'. Below it, a sidebar has 'Create', 'Import', and a download icon. The main area is titled 'Document processing automations (1)' and shows a single entry: 'Clandis Baker Project' last edited on 02/23/2023. Below this, there are categories: 'Published automation services' (with an arrow), 'Decision' (with 'Document processing' selected and an arrow), 'Workflow' (with an arrow), and 'External' (with an arrow). A 'Learn more' link is also present.

2. Select <your project name> Click open

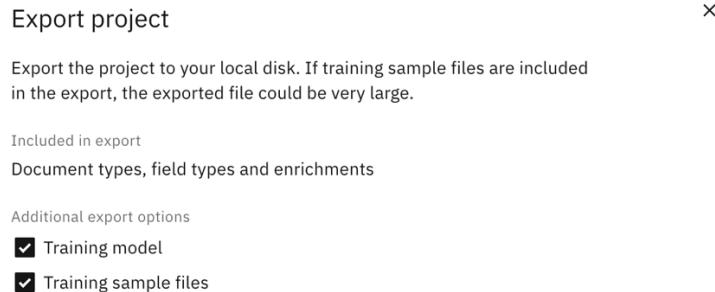
This screenshot is similar to the previous one, showing the 'Business automations' screen. The 'Document processing' category is still selected under 'Decision'. However, the 'Clandis Baker Project' entry now has a large blue 'Open' button with a red arrow pointing to it, indicating it has been selected.

3. From the Main screen select the Configure tab

The screenshot shows the 'Clandis Baker Project' configuration screen. The top navigation bar includes 'IBM Cloud Pak | Administration', 'Business automations /', and the project name 'Clandis Baker Project'. On the right, there are 'Share' and 'Version / Deploy' sections. The main area has tabs for 'Build', 'Enrich', and 'Configure', with 'Configure' being the active tab. Under 'Configure', there are two main sections: 'Import / Export ontology' (with 'Language settings' and 'Git server configuration' options) and 'Export project' (with an 'Export project' button). Below that is an 'Import project' section with a note about deployment status and an 'Import project' button. A status bar at the bottom indicates 'Last shared | 2 hours ago' and 'Latest version | v2 | 2 hours ago'.

_4. Select Export Project

_5. On Export Project window **check Training Module and Training Sample files**



_6. Click on OK

_7. A project-export-<date-time>.zip will be download via browser to local machine.

Appendix A - Troubleshooting

TechZone Pending Status taking Long Time

Operator shows Pending status in a namespace – OLM know issue.

An operator fails to install and continuously shows Pending status.

For fix visit below link.

<https://www.ibm.com/docs/en/cpfs?topic=ii-operator-shows-pending-status-in-namespace-olm-known-issue>

Can't find user/password in Daffy

If your deployment has FAIL when looking into getting username and password then your environment is not working.

```
#####
#          Daffy Options          #
#####
Please use this tool and select what option you would like to retrieve more info on.
With this menu you can get your OpenShift Console URL, id/passwords and status.
You can also get your CP4BA Console URL, id/passwords and status info.

1) OpenShift
2) Services
3) ExitMenu
#? 2
CP4BA Services Menu:
1) Console
2) Status
3) Back
#? 1
#####
                    Running daffy service process v2023-01-11
                    Log File - /data/daffy/log/ocpinstall/cp4ba/service.sh-2023-03-05-10-47.log
#####
Start time : Sun Mar  5 10:47:01 EST 2023

Checking OS before continuing on
#####
Linux is being used (Red Hat Enterprise Linux 8.7 (Ootpa))

Login via oc(ocpadmin)
#####
oc login https://api.ocpinstall.gym.lan:6443 -u ocpadmin -p ***** --insecure-skip-tls-verify
admin user - ocpadmin

Validate OCP Access
#####
✓ PASSED Access to cluster via oc command

Validate CP4BA version info
#####
✓ PASSED Valid version CPBA_VERSION=22.0.2

CP4BA Configuration
#####
Daffy Version           : v2023-01-11
Bastion OS              : rhel - 8.7
Platform Install Type   : vsphere-ipi
OpenShift Cluster Name  : ocpinstall
OpenShift Version        : 4.10.36
CP4BA Version           : 22.0.2
Project/Namespace       : cp4ba-starter
Zen Version              : 4.8.0
Message 1                : Running reconciliation
Message 2                : Prerequisites execution done.
Message 3                : FAIL - prerequisites Deployment failed ←
Message 4                :
Deployment Service       : Starter docprocessing
Config Map Dump          : /data/daffy/log/ocpinstall/cp4ba/icp4adeploy-cp4ba-access-info.yaml

Console Automation Document Processing
#####
```

Environment verification

Once you have reserved a cluster in IBM TechZone, it is first ****Scheduled**** for provisioning. After a while it moves into status ****Provisioning****, and after some time finally becomes

Ready.

At that time, you'll also get an email that your cluster is Ready, but this only means that the Red Hat OpenShift part is now available. Once the cluster is Ready, the deployment of the CP4BA Starter pattern will automatically be performed. Therefore, you must wait until not only the OCP cluster has been provisioned but also until CP4BA Starter pattern has been completely deployed.

*****Combined this may take several hours (~5-6 hours).*****

At the moment, there is a known Red Hat OpenShift bug that can intermittently block the successful deployment of CP4BA Starter pattern. To identify that your TechZone provisioned environment has hit this issue, **please check about one hour after the cluster has become ready** if your cluster is affected by this bug.

For this, please perform the following steps:

- Open the *OpenShift web console* in a browser.
- In the left-hand side navigator go to *Operators* -> *Installed Operators*.
- Make sure the *project scope* is set to *All Projects*.
- Verify that *all Operators* show in the column with *Status* the value *Succeeded*.
- If there are one or multiple Operators *NOT with Status 'Succeeded'* (for example in Status 'Failed', 'Unknown', or 'Cannot update'), your environment is affected by the mentioned bug and applying a manual workaround is required. For this, please reach out for [Support](#support).
- Once all Operators show in column *Status 'Succeeded'*, you can proceed with the next prerequisite.

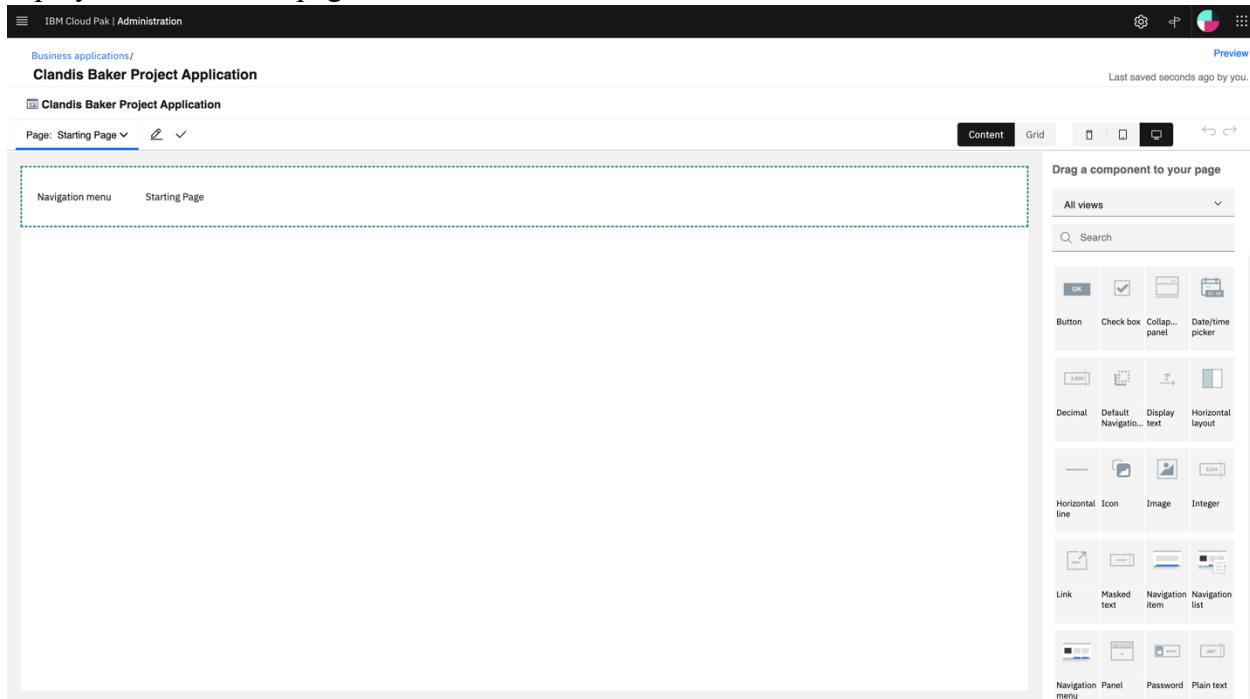
To verify that your CP4BA cluster is completely deployed:

- Open the **OpenShift web console** in a browser.
- Click on **Workloads -> ConfigMaps** on the left-hand side navigator.
- Type '**access-info**' in the field next to 'Name'.

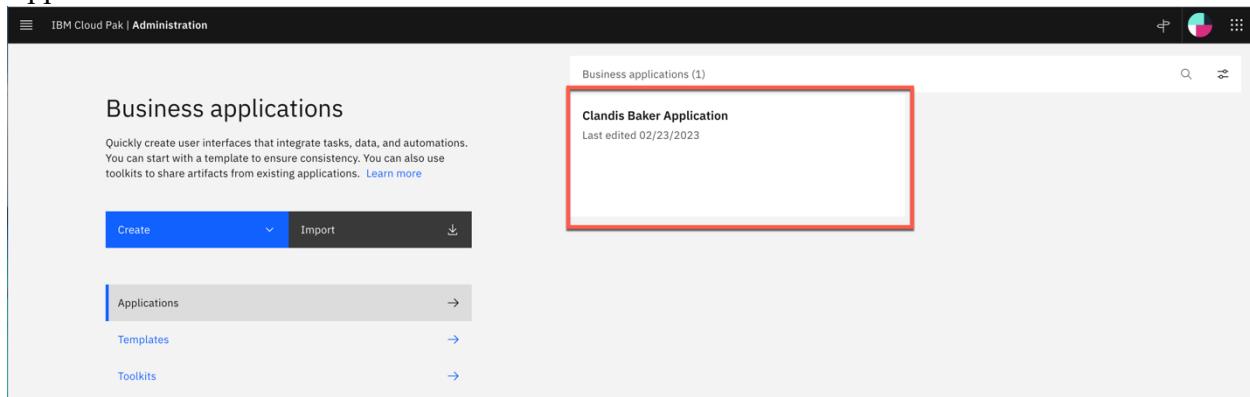
If the ConfigMap '**icp4adeploy-cp4ba-access-info**' is shown, your CP4BA cluster is deployed.

APPLICATION BLANK

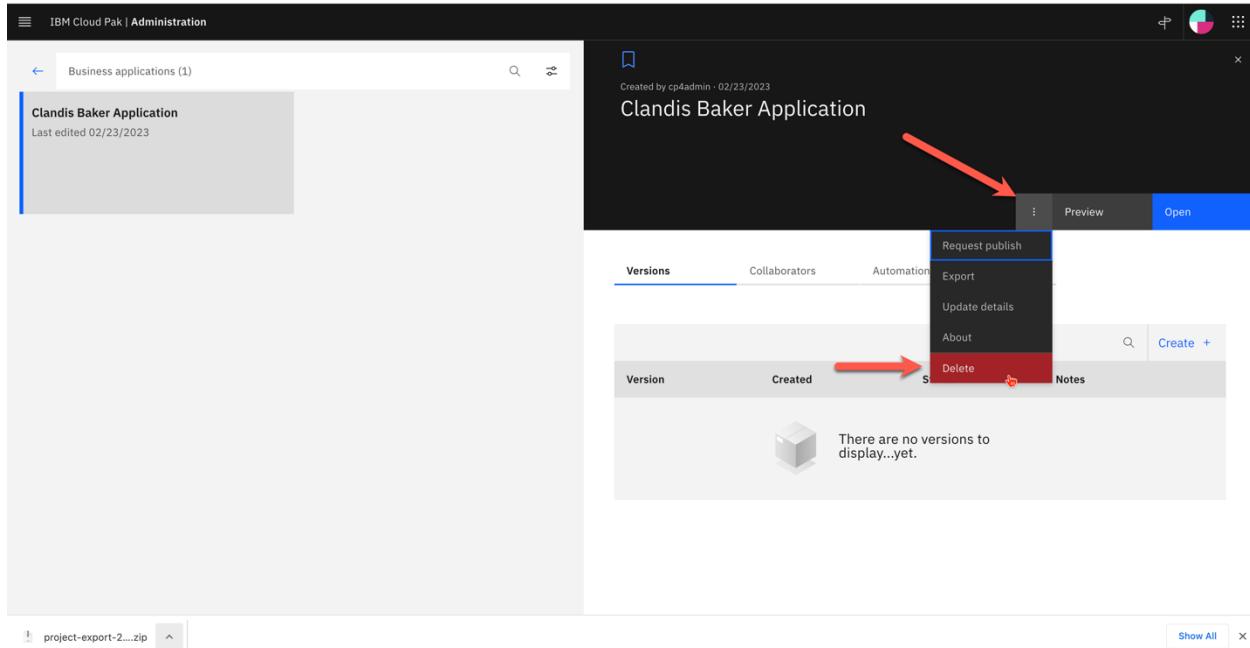
During creating of Business Application setup, sometimes on first time after project has been deployed. The Starter page is blank.



If this happens delete the application and try again. To delete the application, Click on the Application tile



Then Click on the 3 dots and Select Delete



Connection issue with Workstation to Cloud.

If issues with connection from workstation to cloud after it's been working. Reboot your workstation.

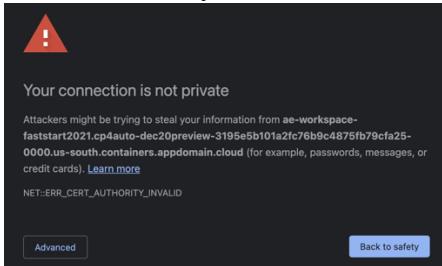
OPENING AN INCOGNITO WINDOW

When you open a new incognito window, you will need to accept certificates before logging in to ADP. Customers shouldn't have this issue because they will have their own certificates instead of the self-signed certificates used in this environment.

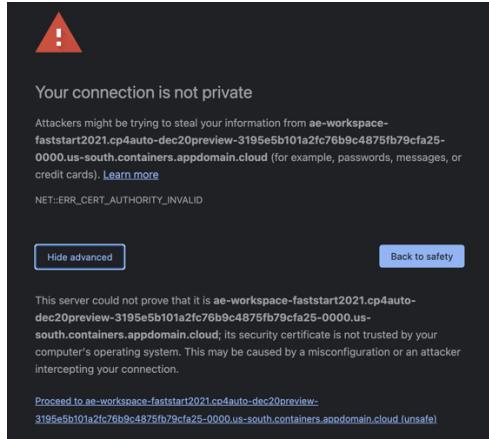
In your incognito window, go to the following URLs located in this Box:

Open the Generate Security Tokens Box note and click all 3 of the links listed. This will reset the self-signed security certificates.

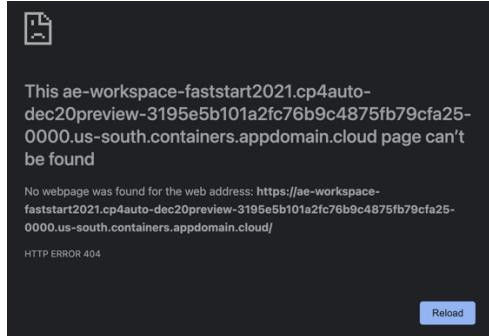
For each URL, your browser window will show a message like this:



Click Advanced, and the browser window will look something like this:



Click the “Proceed to...” link. You’ll see a message like this in your browser window:



Ignore the error and proceed to the next link.

After doing this for each of the URLs above, log in to BAStudio

Appendix B - BAW & ADP Integration Sample

<https://github.com/IBM/baw-adp-integration-sample>