

Automation Document Processing Lab

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.....	7
2 Getting started	8
2.1 IBM TechZone – Reserve the environment.....	9
2.1.1 Credentials.....	Error! Bookmark not defined.
2.2 Set up WireGuard VPN	11
2.3 Open your IBM Cloud Environment	13
3 Lab Overview.....	18
3.1 How does ADP work?	18
4 Create Document Processing Project	20
4.1 Reviewing the interface.....	25
4.1.1 Build Tab	25
4.1.2 Enrich Tab	26
4.1.3 Configure Tab.....	27
5 Configure a Wage and Tax document type.	30
5.1 Create Wage and Tax document type.....	30
5.2 Create Field	32
5.3 Create the Employee Name Address field.....	34
5.4 Create Employee Social Security Number Field.....	35
6 Document Types and Samples Overview.....	39
6.1 Categorize documents.	40
7 Train classification.....	47
7.1 How do I improve my results?	51
8 Data extraction.....	53
8.1 Correcting extracted values	55
8.2 Train extraction model.....	60
9 Data standardization	61
10 Version and deploy your project	63
11 Application designer.....	65
11.1 Create your Runtime Application.	65
11.2 Upload documents for processing	70
11.3 Correct any classification errors.....	72
11.4 Correct extraction issues.....	74
12 Export Import Project.	82
Appendix A - Troubleshooting	84
TechZone Pending Status taking Long Time	84
Can't find user/password in Daffy	84
APPLICATION BLANK	87
Connection issue with Workstation to Cloud.....	88
OPENING AN INCOGNITO WINDOW.....	88

Appendix B - BAW & ADP Integration Sample.....	90
https://github.com/IBM/baw-adp-integration-sample	90

1. Overview

1.1 Getting HELP during the lab

- For internal IBM, another good resource is 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 Icons

The following symbols appear in this document at places where additional guidance is available.

Icon	Purpose	Explanation
	Important!	This symbol calls attention to a particular step or command. For example, it might alert you to type a command carefully because it is case sensitive.
	Information	This symbol indicates information that might not be necessary to complete a step but is helpful or good to know.
	Trouble-shooting	This symbol indicates that you can fix a specific problem by completing the associated troubleshooting information.

•

1.3 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.

1.4 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.

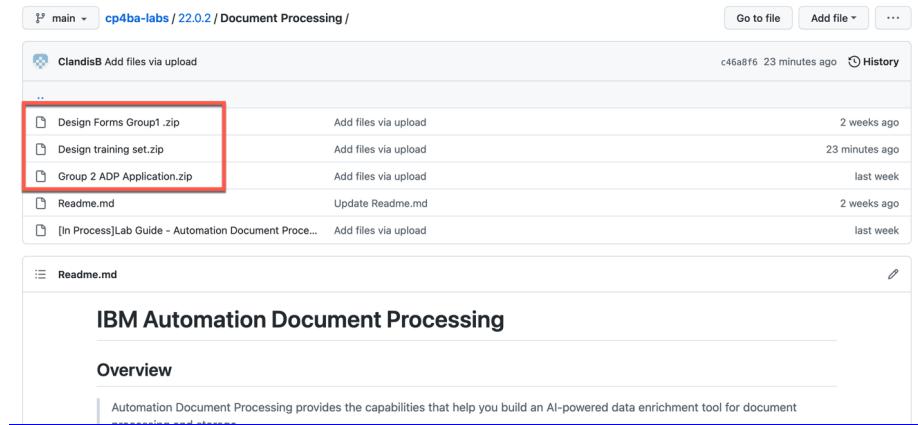
You can extract text, check boxes, forms, tables, barcodes, signature detection and even free text. With no or low code options, you can create an application that processes documents, extracts data, flags issues, and stores your documents and data. And the data enrichment capabilities ensure that the 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. We will be using these sample documents during the labs. You can find them here:

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



main · cp4ba-labs / 22.0.2 / Document Processing

ClandisB Add files via upload c46a8f6 23 minutes ago History

Design Forms Group1.zip Add files via upload 2 weeks ago

Design training set.zip Add files via upload 23 minutes ago

Group 2 ADP Application.zip Add files via upload last week

Readme.md Update Readme.md 2 weeks ago

[In Process]Lab Guide - Automation Document Proce... Add files via upload last week

Readme.md

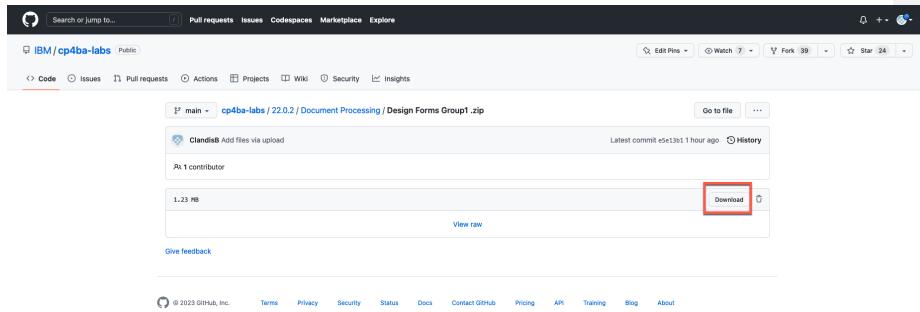
IBM Automation Document Processing

Overview

Automation Document Processing provides the capabilities that help you build an AI-powered data enrichment tool for document

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

_2. Then Click on Download



Search or jump to... Pull requests Issues Codespaces Marketplace Explore

IBM / cp4ba-labs Public

Code Issues Pull requests Actions Projects Wiki Security Insights

main · cp4ba-labs / 22.0.2 / Document Processing / Design Forms Group1 .zip

ClandisB Add files via upload Latest commit e5c13b1 1 hour ago History

An 1 contributor

1.23 MB

Download

View raw

Give feedback

_3. Repeat above steps “Group 2 ADP Application.zip” and “Design training set.zip”

You will notice the images are in various unique folders that will be referenced specifically in the different labs later. Please keep them in their proper folders.

2.1 IBM TechZone – Overview

What is IBM TechZone?

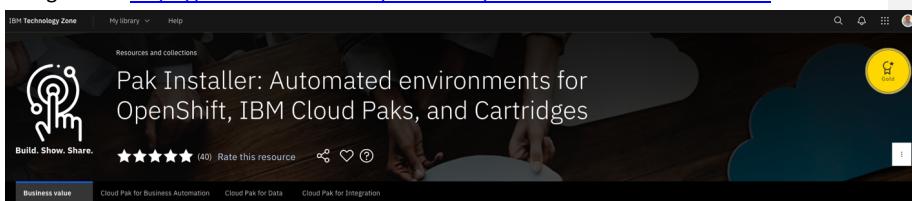
IBM Technology Zone (techzone.ibm.com) enables IBM teams and IBM Business Partners to provision technical “Show Me” live environments, Proof-of-Technologies, prototypes, and Minimum Viable Prototypes, which can be customized, shared with peers and clients to experience IBM Technology.

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

The TechZone leverages DAFFY. DAFFY is Deployment Automation Framework For You. The DAFFY installer tool has been renamed to Pak Installer. This tool will do all the heavy lifting of the OpenShift and IBM Cloud Pak installs. The National Market Top Team created Pak Installer to assist the technical sales teams with the progression of IBM Cloud Pak opportunities. The goal is to provide the technical sales with a set of (easy to use) scripts that will aid in the installation of OpenShift and the IBM Cloud Pak's. For more information on DAFFY/Pak Installer please look at: <https://ibm.github.io/daffy/>

2.1.1 Reserve Environment

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



The screenshot shows the IBM TechZone interface. At the top, there's a navigation bar with 'IBM Technology Zone', 'My library', and 'Help'. Below the header, a banner for 'Pak Installer: Automated environments for OpenShift, IBM Cloud Paks, and Cartridges' is displayed, featuring a 5-star rating and a 'Rate this resource' button. The main content area shows a large image of a person interacting with a white tablet or screen. Below the image, 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.' At the bottom of the page, there are sections for 'Authors' (listing Kyle Dawson), 'Resources' (listing 'Cloud Pak for Business Automation'), and 'Comments'.

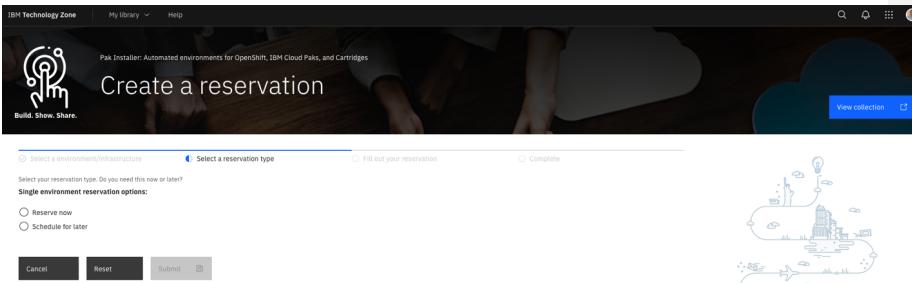


Note: This environment is built with Daffy by Kyle Dawson with the latest releases. This environment can also be used at a customer site with same tool and framework of Daffy.

- _2. Click Cloud Pak for Business Automation tab and scroll down to the “Cloud Pak for Business Automation 22.0.2 – VMWare tile.
_3. Click on Reserve

Automation Document Processing Lab

- _4. On Create a reservation screen **select option** for when to start



- _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

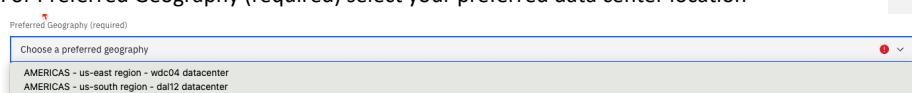
Standard proof of concept; Need an environment for a standard product use case.

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

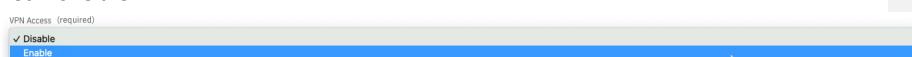
Testing: Need to test a specific function, configuration, or customization.

- _6. For this lab **Select 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



- _8. For VPN Access **choose Enable**. You will be using a VPN to connect from desktop to the TechZone tile

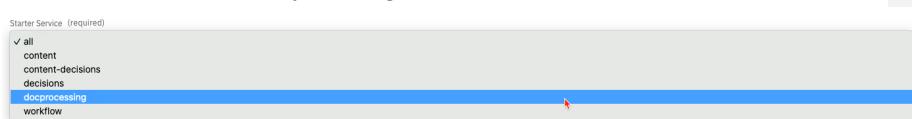


Make sure to pick enable otherwise you'll have to start all over with deployment.

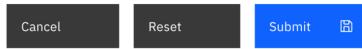
- _9. In Cloud Pak for Business Automation Version Pick 22.0.2. (if not already chosen for you)

- _10. In Cloud Pak for business Automation IFix pick IF002 (if not already chosen for you)

- _11. For Starter Service **choose docprocessing**



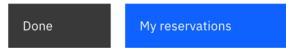
_12. Click Submit



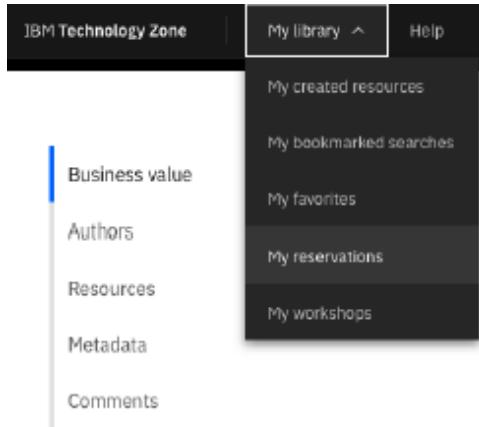
Upon receiving the Your environment is ready email, please allow up to 1 hour for the start-up services to fully complete. If after receiving email and a few hours have passed and your environment is not up, check [Appendix A – Trouble Shooting](#) for possible fix.

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.

_13. Click My reservations



_14. 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>

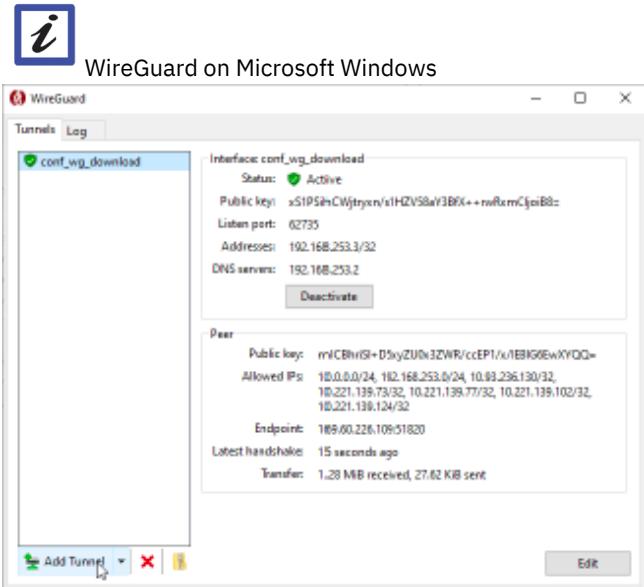
2.2 Set up WireGuard VPN

_4. Open your reservation tile and scroll to bottom.

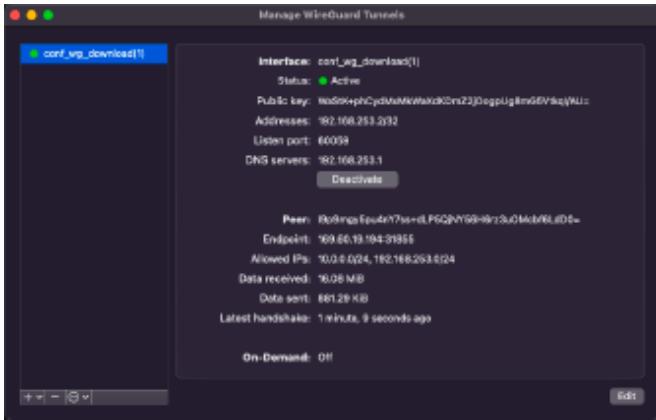
- _5. Click Download WireGuard VPN config button to download conf_wg_download.conf to your local workstation

Download Wireguard VPN config 

- _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.



Automation Document Processing Lab



2.3 Open your IBM Cloud Environment

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

A screenshot of the IBM Technology Zone interface. It shows a reservation for "Cloud Pak for Business Automation 22.0.1/22.0.2 - VMWare (Powered by Pak Installer)". The reservation was created on Feb 26, 2022, at 7:45 AM and will expire on Feb 27, 2022, at 7:45 AM. The status is Ready. On the right, there is a clock icon and a calendar for August. Below the reservation details, there is a "Desktop" section with a "Purpose" field containing "Open your IBM Cloud environment". It also includes "Notes" and "Environment" fields. A "Shared Reservation" note states that full desktop access can be connected via a provided URL.

- _2. Let's get the username and password created by DAFFY. Expand OCP Gym under All Connections

ALL CONNECTIONS

- ⊖ OCP Gym
- ⊖ gym-06000180cg-ar09hjws
 - _ CloudPak Information
 - Remote Red Hat Linux Desktop
 - _ SSH Command Line

_3. Select CloudPak Information

ALL CONNECTIONS

- ⊖ OCP Gym
- ⊖ gym-06000180cg-ar09hjws
 - _ CloudPak Information
 - Remote Red Hat Linux Desktop
 - _ SSH Command Line

_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.

```
Validate CP4BA version info
#####
✓ PASSED  Valid version CPBA_VERSION=22.0.2
✓ PASSED  Valid IFIX CP4BA_IFIX=IF002

CP4BA Service Status
#####
Daffy Version           : v2023-03-09
Bastion OS              : rhel - 8.7
Platform Install Type   : vsphere-ipi
OpenShift Cluster Name  : ocpinstall
OpenShift Version        : 4.10.36
CP4BA Version           : 22.0.2 IF002
Project/Namespace       : cp4ba-starter
Zen Version              : 4.8.1
Message 1                : Running reconciliation
Message 2                : Prerequisites execution done.
Message 3                :
Message 4                :
Deployment Service      : Starter docprocessing
Config Map Dump          : /data/daffy/log/ocpininstall/cp4ba/icp4adeploy-cp4ba-access-info

Console Automation Document Processing
#####
Cloud Pak Business 3     : https://cp4ba-starter.apps.ocpininstall.gym.lan
Cloud Pak Admin Username  : cp4admin
Cloud Pak Admin Password  : Tm1WRtxkUI1bv2drooMF

#####
End Time: Mon Mar 20 11:43:53 EDT 2023
CP4BA Service Completed in 10 seconds
#####

CP4BA Services Menu:
Please select 1,2 or 3
#####
1) Console
```



Note: Controls for copy and paste in guacamole.

For Mac users:

CONTROL_OPTION_SHIFT

For Windows users:

CTRL_ALT_SHIFT



If your screen shows FAIL then it's not ready just yet and wait a bit longer.

```
# 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
✓ PASSED Valid IFIX CP4BA_IFIX=IF002

CP4BA Service Status
#####
Daffy Version : v2023-03-09
Bastion OS : rhel - 8.7
Platform Install Type : vsphere-ipi
OpenShift Cluster Name : ocpinstall
OpenShift Version : 4.10.36
CP4BA Version : 22.0.2 IF002
Project/Namespace : cp4ba-starter
Zen Version : 4.8.1
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
#####
Cloud Pak Dashboard : 
```

- _7. Back on your Reservation tile **copy or click the link Cloud Pak Dashboard URL** and paste 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/Accept the Risk and Continue. This may occur twice.



Warning: Potential Security Risk Ahead

Firefox detected a potential security threat and did not continue to **cpd-cp4ba-starter.apps.ocpinstall.gym.lan**. If you visit this site, attackers could try to steal information like your passwords, emails, or credit card details.

What can you do about it?

The issue is most likely with the website, and there is nothing you can do to resolve it.

If you are on a corporate network or using antivirus software, you can reach out to the support teams for assistance. You can also notify the website's administrator about the problem.

[Learn more...](#)

[Go Back \(Recommended\)](#)

[Advanced...](#)

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

Commented [CB1]: See you can add cp4admin

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 an automate 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 by making the capabilities and artifacts available for integration into an application and into the destination repository.

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.

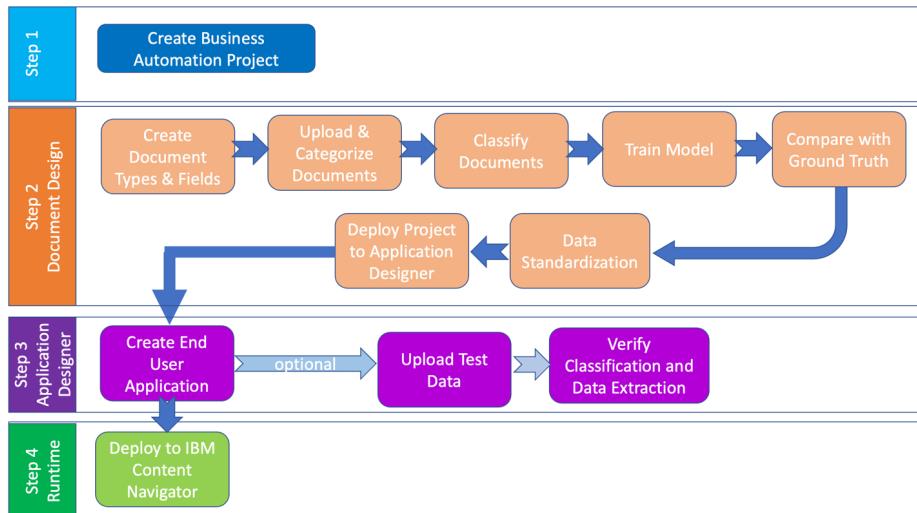
When you deploy your document processing application, you connect it to a content repository that manages the document types and the extracted data for each document. The solution is fully integrated with IBM FileNet® Content Manager, simplifying document and data storage by applying your existing filing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

End result

Automation Document Processing Lab

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.



Step1 – 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. To get more information on creating/using the Business Automation Application (BAA) look at the SWAT Jam Lab for BAA.

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

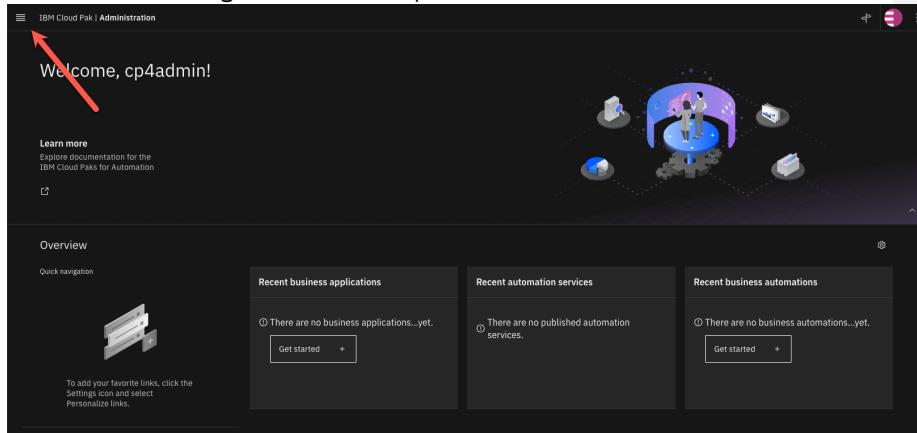
Step 1

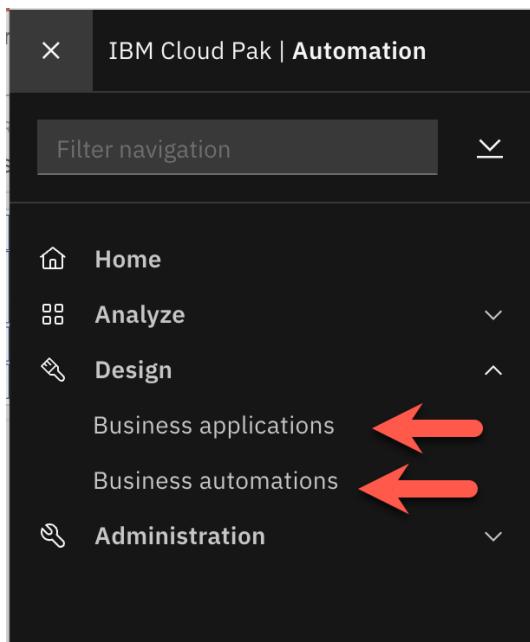
Create Business Automation Project

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.

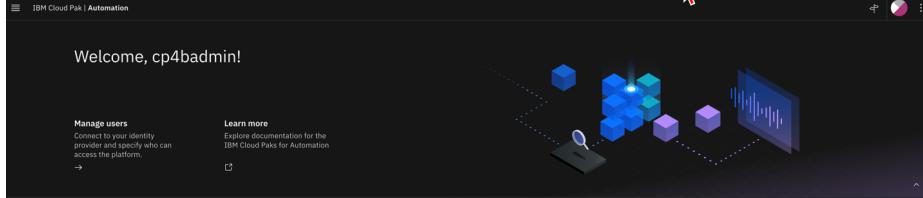
Within the Business Automations you can create or reuse automations. An automation is a collection of artifacts that fulfills a business purpose. You can publish some automation artifacts as automation services that you can call and reuse in a consistent way. Also in Business Automation, you use the **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.

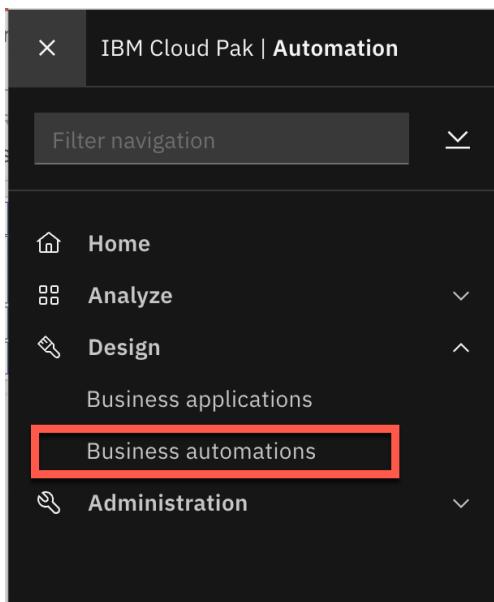
Within **Business Applications** you can quickly create user interfaces that integrate tasks, data, and automations. You can start with a template to ensure consistency. You can also use toolkits to share artifacts from existing applications.

We will start with the Business Automations.
Once logged in to the IBM Automation Server, you should see the Welcome screen.

Automation Document Processing Lab



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 navigation bar, 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, there are two main buttons: 'Create' (highlighted in blue) and 'Import'. Underneath these buttons, a list of service categories is shown, each with a right-pointing arrow: 'Published automation services', 'Decision', 'Document processing', 'Workflow', and 'External'. The 'Document processing' category is highlighted with a blue border.

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

The screenshot shows the same 'Business automations' page as the previous one, but with a red arrow pointing to the 'Create' button in the top navigation bar. A dropdown menu has opened from the 'Create' button, listing several options: 'Decision automations', 'Document processing automations' (which is highlighted with a red box), 'Workflow', and 'External'. Below this dropdown, the list of service categories is identical to the first screenshot: 'Published automation services', 'Decision', 'Document processing', 'Workflow', and 'External'. The 'Document processing' category is also highlighted with a blue border in this view.

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

The screenshot shows a dialog box titled 'Create a document processing automation'. It has two input fields: 'Name' containing 'User01_CEB' and 'Purpose (optional)' containing 'My project for user01'. At the bottom right are 'Cancel' and 'Create' buttons, with 'Create' being highlighted in blue.

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

Automation Document Processing Lab

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 icon. Below it, a breadcrumb path shows 'Business automations / User01_CEB'. On the right, there are 'Share' and 'Version / Deploy' buttons. 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), 'Classification model' (3 types trained, 100% accuracy), 'Extraction model' (3 types trained, 95% accuracy), 'Data standardization' (Not ready, Start button), and 'Document retention' (3 types reviewed). Each section includes a status indicator (Ready or Not ready) and an 'Open' button.

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).

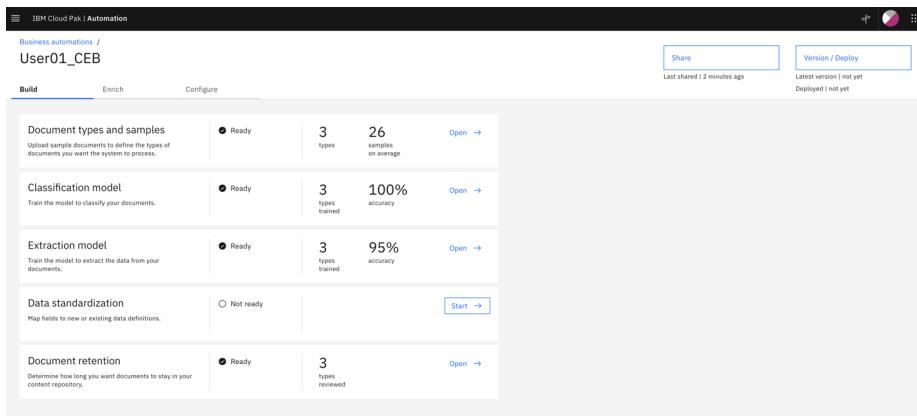
Automation Document Processing Lab

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.



The screenshot shows the IBM Cloud Pak | Automation interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a search bar. Below it, a breadcrumb trail shows 'Business automations / User01_CEB'. On the right, there are 'Share' and 'Version / Deploy' buttons, with status information: 'Last shared 2 minutes ago', 'Latest version | not yet Deployed | not yet'.

The main area has three tabs: 'Build', 'Enrich' (which is selected), and 'Configure'. The 'Enrich' tab displays several sections with status indicators (Ready or Not ready) and counts (e.g., 3 types, 26 samples, 100% accuracy). Buttons like 'Open' and 'Start' are present. A 'Document retention' section is also shown.

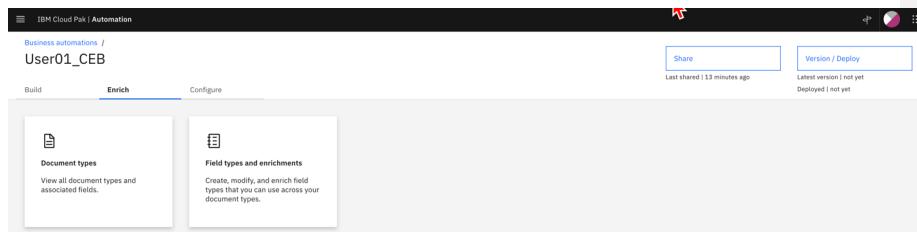
Section	Status	Count	Details
Document types and samples	Ready	3 types	26 samples on average
Classification model	Ready	3 types trained	100% accuracy
Extraction model	Ready	3 types trained	95% accuracy
Data standardization	Not ready		
Document retention	Ready	3 types reviewed	

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.

Automation Document Processing Lab



- _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.

Business automations / Clandis Baker Project /

Field types and enrichments

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

Automation Document Processing Lab

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 of the 'Clandis Baker Project' in the 'Business automations' section of IBM Cloud Pak. The 'Import / Export ontology' section includes 'Language settings' and 'Git server configuration'. The 'Export project' section has a button to 'Export project'. The 'Import project' section has a note about deployment status and a 'Import project' button.

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.

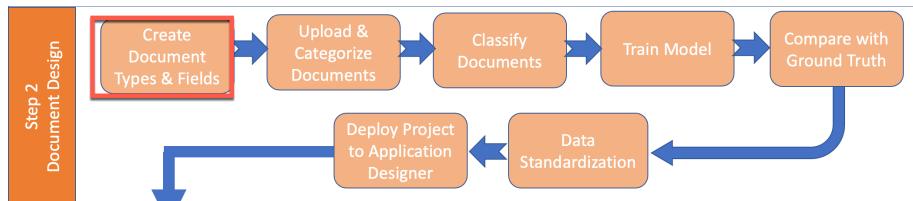
Automation Document Processing Lab

The screenshot shows the 'Language settings' section of the 'Configure' tab for the 'Clandis Baker Project'. It includes fields for 'Extraction language' (set to English), 'Display name language' (set to English (en) (default)), and a 'Project locale' dropdown. Buttons for 'Share', 'Version / Deploy', and 'Save' are visible.

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.

The screenshot shows the 'Git server configuration' section of the 'Configure' tab for the 'Clandis Baker Project'. It includes fields for 'Git vendor' (set to Gitea), 'Git server organization URL' (set to <https://cp4adeploy-gitea-svc:3000/content-designer>), 'Git server REST API URL' (set to <https://cp4adeploy-gitea-svc:3000/api/v1>), 'Username' (set to gt), 'Type of credentials' (set to API key), and a 'Credentials' input field. Buttons for 'Test' and 'Save' are visible.

5 Configure a Wage and Tax document type.

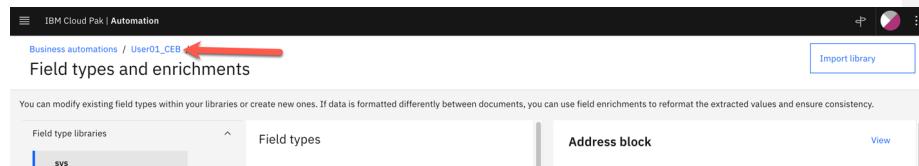


Before we use the guided configuration, you will configure some additional document types and fields used to extract data prior to uploading sample documents.

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

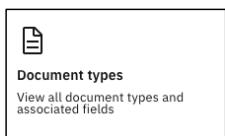
5.1 Create Wage and Tax document type.

- _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



- _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 X

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

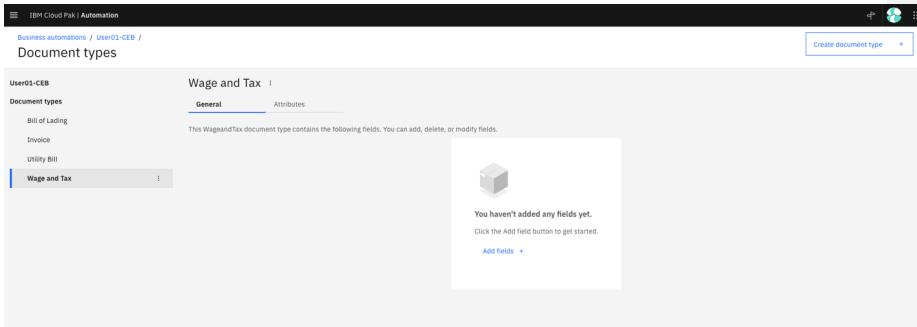


Note: Notice the option for “Fixed-format document type”. If your form is static in nature or has a fixed structure that does not change, select this option so you will not have to provide as many samples. In our use case Wage and Tax documents have a variety of formats and are not static.

_6. Click the ADD button.

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

Automation Document Processing Lab



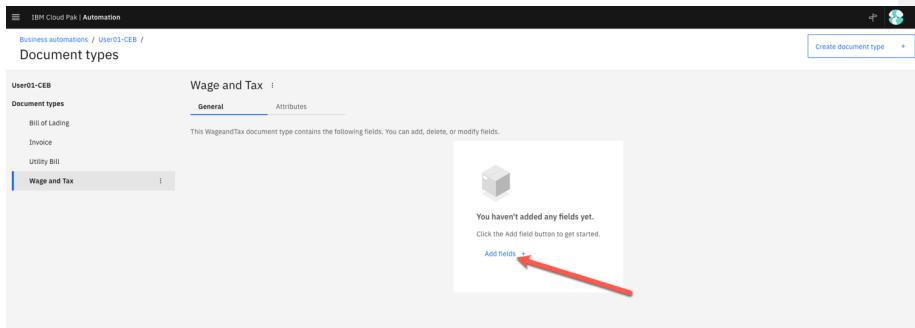
The screenshot shows the 'Document types' section of the IBM Cloud Pak | Automation interface. On the left, there's a sidebar with 'User01-CEB' and 'Document types' sections, listing 'Bill of Lading', 'Invoice', 'Utility Bill', and 'Wage and Tax'. The 'Wage and Tax' item is highlighted with a blue border. The main area has tabs for 'General' and 'Attributes', with 'General' being the active tab. A central panel displays the message: 'You haven't added any fields yet. Click the Add field button to get started.' Below this message is a blue 'Add fields' button. A red arrow points to this 'Add fields' button.

_7. Select your **Wage and Tax doc 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 configuration page for the 'Wage and Tax' document type. It has the same layout as the previous screenshot, with the 'General' tab active. The central panel still says 'You haven't added any fields yet.' and has the 'Add fields' button. A prominent red arrow points directly at the 'Add fields' button.

_2. Enter the following values under the GENERAL Settings header

Automation Document Processing Lab

The screenshot shows the 'Create field' dialog in the IBM Cloud Pak | Automation interface. The document type selected is 'Purchase Orders'. The 'General' tab is active. In the 'Display name' field, the value 'Ex. Employee's name, Le nom de l'employé' is entered, which is highlighted with a red border and has a red error dot next to it. The 'Description (optional)' field is empty. Under 'Symbolic name', the value 'Enter a name' is shown. The 'Field type' dropdown is set to 'sys:String'. In the 'Aliases' section, there is a text input field with 'Enter an alternative name' and a note 'Enter an alternative name and press the "Enter" key'. There are two checkboxes at the bottom: 'This field is required' (unchecked) and 'This field contains sensitive information' (unchecked). Navigation buttons 'Cancel' and 'Next' are visible at the top right.

- **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:

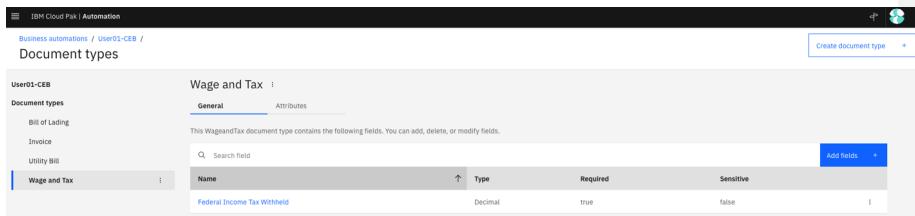
The screenshot shows the 'Create field' dialog in the IBM Cloud Pak | Automation interface. The document type selected is 'Wage and Tax'. The 'General' tab is active. In the 'Display name' field, the value 'Federal Income Tax Withheld' is entered. In the 'Symbolic name' field, the value 'FederalIncomeTaxWithheld' is entered. The 'Field type' dropdown is set to 'sys:Decimal'. In the 'Aliases' section, there is a text input field with 'Enter an alternative name' and a note 'Enter an alternative name and press the "Enter" key'. Below the input field, the aliases '2 Federal income tax withheld' and '2. Federal income tax' are listed, separated by a plus sign. Navigation buttons 'Cancel' and 'Next' are visible at the top right.

_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.



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

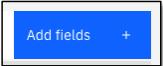
Automation Document Processing Lab

The screenshot shows the 'Employee name and address' field configuration page. At the top, it says 'Document type: Form W2'. Below that, there are tabs for 'General', 'Field patterns', 'Value settings', and 'Subfields'. Under 'General Settings', the 'Display name' is 'Employee name and address' and the 'Symbolic name' is 'Employee name and address'. The 'Field type' is 'sys:Address information'. There are checkboxes for 'This field is required' (checked) and 'This field contains sensitive information' (unchecked). On the right, there are sections for 'Description (optional)' and 'Aliases'. The 'Description' field is empty, and the 'Aliases' field contains 'Employee's first name and initial Last name Suffix', 'Employee's name, address, and ZIP code', and 'Employee's name, address, and ZIP code'. Buttons for 'Cancel' and 'Next' are at the bottom right.

- _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**

Automation Document Processing Lab

Your screen should now look like the image below:

The screenshot shows the 'Employee Social Security Number' field configuration in the 'General' tab. The 'Display name' is 'Employee Social Security Number' and the 'Symbolic name' is 'EmployeeSocialSecurityNumber'. The 'Field type' is set to 'sys:Numeric'. There are two checkboxes: one checked for 'This field is required' and another unchecked for 'This field contains sensitive information'. In the 'Aliases' section, there is an entry 'Employee's social security number'.

_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

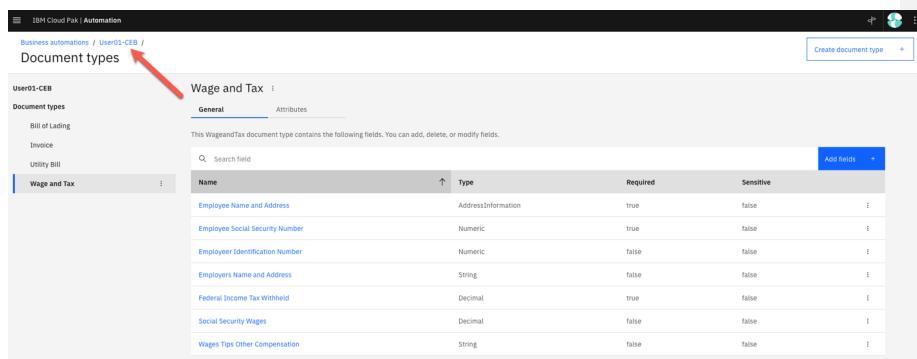
Reference for various field types:



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>

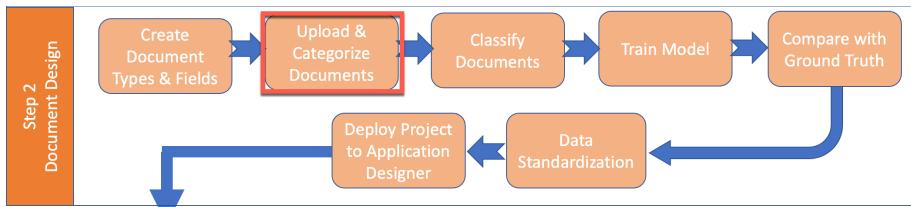
- _6. Click on the <name of your project> in the breadcrumb link in the top left of your screen. In the following example the name of the project is <User01_CEB>.



The screenshot shows the 'Document types' section of the IBM Cloud Pak for Automation interface. The breadcrumb navigation bar at the top indicates the path: 'IBM Cloud Pak | Automation' > 'Business automations' > 'User01-CEB' > 'Document types'. A red arrow points to the 'User01-CEB' part of the breadcrumb. The main content area displays a table of fields for a 'Wage and Tax' document type. The table has columns for 'Name', 'Type', 'Required', and 'Sensitive'. The fields listed are: Employee Name and Address (AddressInformation), Employee Social Security Number (Numeric), Employer Identification Number (Numeric), Employer Name and Address (String), Federal Income Tax Withheld (Decimal), Social Security Wages (Decimal), and Wages Tips Other Compensation (String). The 'Required' column shows values like true or false, and the 'Sensitive' column shows values like false or true.

Name	Type	Required	Sensitive
Employee Name and Address	AddressInformation	true	false
Employee Social Security Number	Numeric	true	false
Employer Identification Number	Numeric	false	false
Employer 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.

The screenshot shows the IBM Cloud Pak | Administration interface. At the top, it says "IBM Cloud Pak | Administration" and "Business automations / Clandis Baker Project". Below that, there are tabs for "Build" (which is selected), "Enrich", and "Configure". On the right, there are buttons for "Share" (Last shared | 2 days ago) and "Version / Deploy" (Latest version | not yet Deployed | not yet). The main area has a section titled "Document types and samples" with a sub-instruction: "Upload sample documents to define the types of documents you want the system to process.". This section is highlighted with a red box. Below it are four cards: "Classification model" (Ready, 3 types trained, 100% accuracy), "Extraction model" (Ready, 3 types trained, 97% accuracy), and "Data standardization" (Not ready). There is also a link to "Document types and samples" with a red arrow pointing to the "Document types and samples" section above.

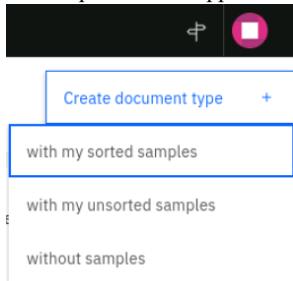
The CATEGORIZE feature analyzes each document and tries to find similarities between them. Based on these similarities, the system will divide the samples into categories for you to review. You can add documents or entire categories into either an existing document class or create new classes as needed.

Let's see what that looks like.

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



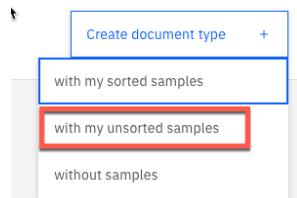
The drop down that appears:



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.

Automation Document Processing Lab

The screenshot shows a web-based application titled 'IBM Cloud Pak | Automation'. The URL in the address bar is 'Business automation / User's CRB / Document types and samples / Create document types'. At the top right are 'Cancel' and 'Categorize' buttons. Below the buttons is a note: 'Upload sample documents that represent the different types of documents you want the system to classify. Include at least 6 samples of each type of document.' There is a search bar labeled 'Search sample documents' and an 'Upload' button with a file icon. A list of documents is displayed in a table format:

Document name
<input type="checkbox"/> Mortgage Agreement1.pdf
<input type="checkbox"/> Mortgage Agreement2.pdf
<input type="checkbox"/> Mortgage Agreement3.pdf
<input type="checkbox"/> Mortgage Agreement4.pdf
<input type="checkbox"/> Mortgage Agreement5.pdf
<input type="checkbox"/> TR_FW2_1001_0000_P5.pdf
<input type="checkbox"/> TR_FW2_2000_0000_P5.pdf
<input type="checkbox"/> TR_FW2_3000_0000_P5.pdf
<input type="checkbox"/> TR_FW2_3001_0000_P5.pdf
<input type="checkbox"/> TR_FW2_4000_0000_P5.pdf
<input type="checkbox"/> UBILLCable_001_1_1.pdf
<input type="checkbox"/> UBILLCable_002_1_1.pdf

At the bottom of the list are pagination controls: 'Items per page' set to 20, '1 - 12 of 12 items', and navigation arrows.

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.

Automation Document Processing Lab

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 screenshots show the 'Create document types' interface in IBM Cloud Pak | Automation. Each screenshot displays a list of sample documents under a specific category. The categories are:

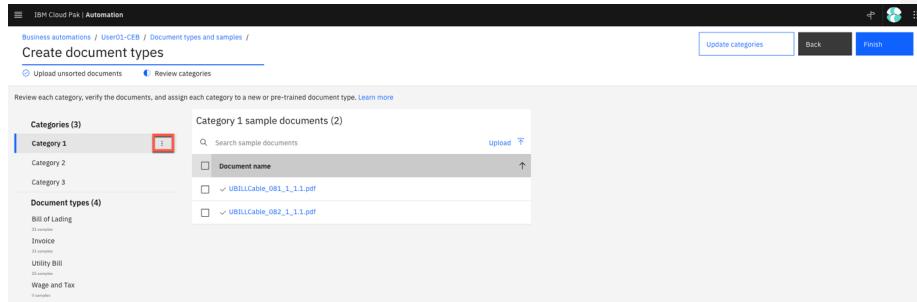
- Category 1 sample documents (2)**: Contains two PDF files: 'UBILLCable_081_1_1.pdf' and 'UBILLCable_082_1_1.pdf'.
- Category 2 sample documents (5)**: Contains five PDF files: 'Mortgage Agreement1.pdf', 'Mortgage Agreement2.pdf', 'Mortgage Agreement3.pdf', 'Mortgage Agreement4.pdf', and 'Mortgage Agreement5.pdf'.
- Category 3 sample documents (5)**: Contains five PDF files: 'TR_FW2_1001_0000_P5.pdf', 'TR_FW2_2000_0000_P5.pdf', 'TR_FW2_3000_0000_P5.pdf', 'TR_FW2_3001_0000_P5.pdf', and 'TR_FW2_4000_0000_P5.pdf'.

Each screenshot includes a sidebar with 'Categories (3)' and 'Document types (4)' sections, and a top navigation bar with 'Update categories', 'Back', and 'Finish' buttons.

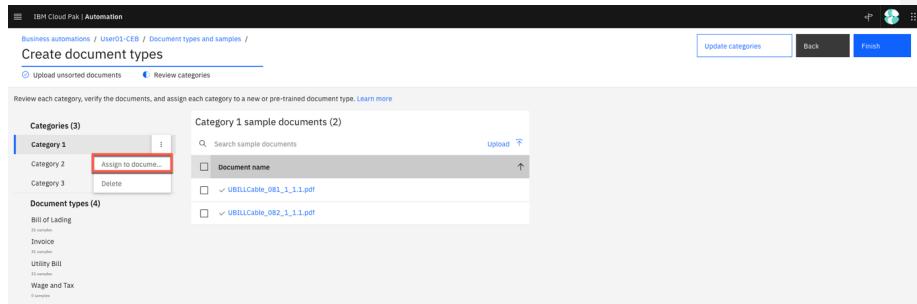
At the time of writing this documentation ADP was able to categorize the sample set into each category. This is not always the case, sometimes document types will be combined into one category, so it's very important to look at each category and verify documents.

Automation Document Processing Lab

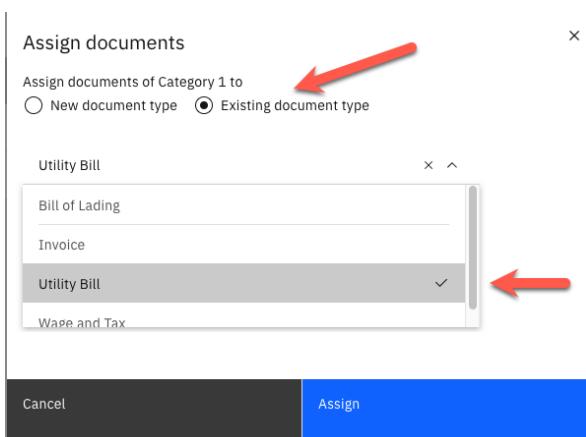
- _7. If all documents within a category are correct as illustrated in the following screen shot, hover over the category name and **Click on the 3 dots** at the end of the category name.



_8. Select ASSIGN TO DOCUMENT TYPE



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



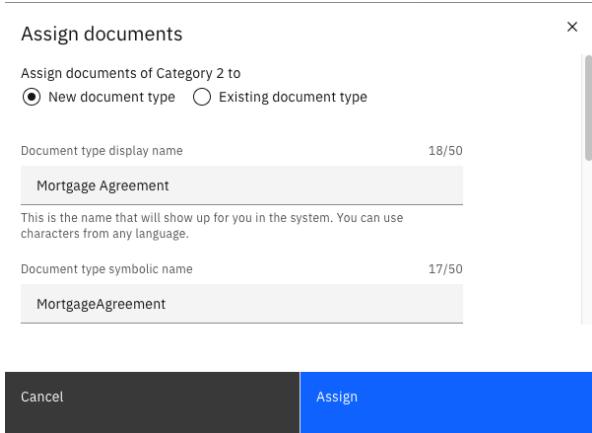
_10. **Click Assign** to close the dialog box

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

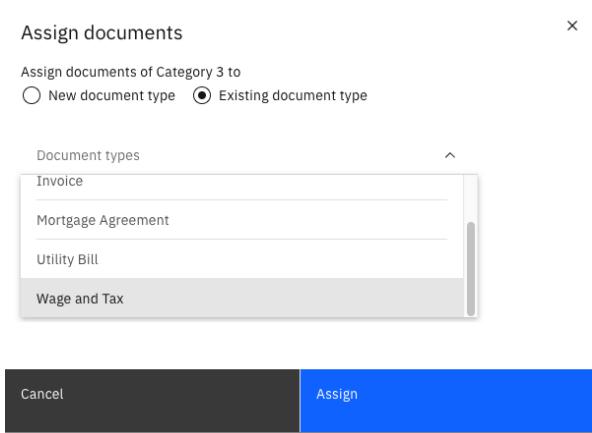
_11. **Select** the next Category 2 and **Click** on the 3 dots and **Select Assign these documents** to a document class.

_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

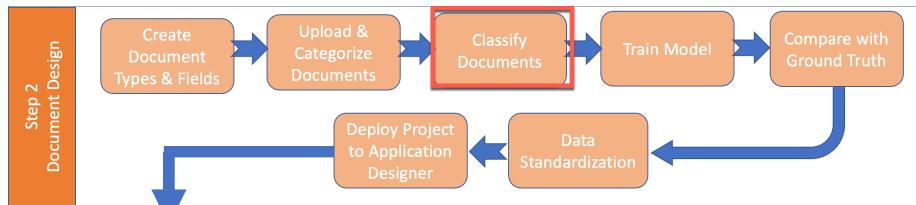


- _14. **Click Assign** to have the system automatically rename and move the category into the Document Types section.
- _15. Now for Category 3, **Click on 3 dots** and Select Assign Document type.
- _16. Select Existing Document Type and Click Wage and Tax from the drop down and then Click on Assign.



- _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

Automation Document Processing Lab

The screenshot shows the IBM Cloud Pak Administration interface for the 'Clandis Baker Project'. The 'Build' tab is active. The 'Document types and samples' section shows 5 types and 20 samples on average. The 'Classification model' section is highlighted with a red box; it shows 3 types trained with 100% accuracy. The 'Extraction model' section shows 3 types trained with 97% accuracy. The 'Data standardization' section shows 'Not ready'. The 'Document retention' section shows 5 types reviewed.

Section	Status	Value
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.

Automation Document Processing Lab

Classification model Accuracy: 84.8%

Document types

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

Mortgage Agreement sample documents (5)

Training set (5)	Test set (0)
100% of total samples	0% of total samples
5 documents	0 documents

Changes were made since you last trained your model. Retrain the model to get updated training results and accuracy.

This document type will not be trained because you have no documents in the test set. Please make sure you have at least 1 document in each 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

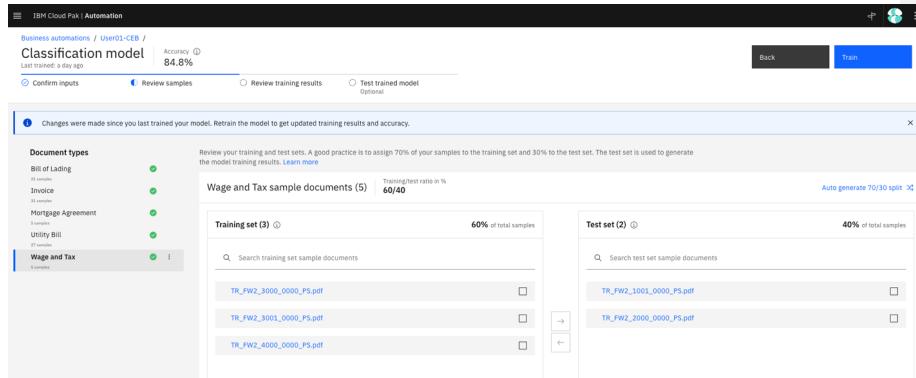
Mortgage Agreement sample documents (5)

Training set (3)	Test set (2)
60% of total samples	40% of total samples
3 documents	2 documents

- _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.

Automation Document Processing Lab



- _6. Click on **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 yielded 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:

Automation Document Processing Lab

The screenshot shows the IBM Cloud Pak for Automation interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a breadcrumb trail: 'Business automation / User@1-CFB / Classification model'. Below this, it says 'Accuracy 96.9%' with a note 'Last trained: 4 minutes ago'. There are four buttons: 'Confirm inputs', 'Review samples', 'Review training results' (which is highlighted in blue), and 'Test trained model optional'. A green modal window appears in the top right corner with the message 'Model trained successfully!' and 'Accuracy has been updated to reflect the latest changes.' In the main content area, there's a message: '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 a list of categories: Bill of Lading, Invoice, Mortgage Agreement, Utility Bill, and Wage and Tax. Under 'Training results', there's a table titled 'These documents are used to test classification. After the classification model is trained, each of these documents is tested to see whether the system can correctly determine the document type.' The table has columns: Document, Classified as, Classification result, and Confidence. It lists several PDF files, all classified as 'Bill of Lading' and marked as 'Correct', with confidence levels ranging from 'High' to 'Medium'.

_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. Low Confidence means we probably need to add more documents to our document class to get better confidence values.

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?

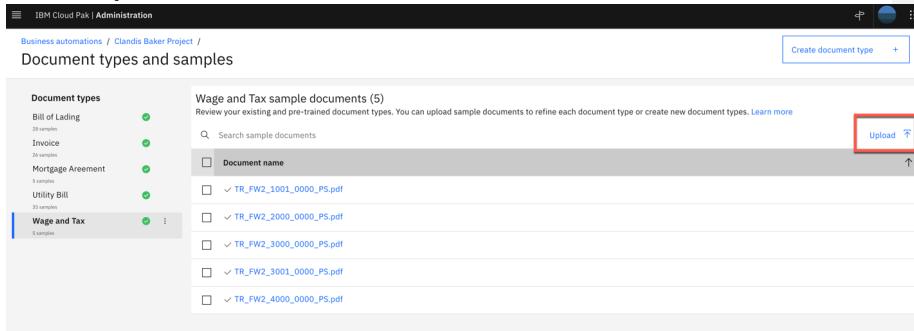
Option 1 – Add more samples.

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.

_1. Click anywhere on Document Types and Samples.

_2. Click on Wage and Tax type.

_3. Click on Upload



_4. From the zip files you downloaded earlier upload all the files from the directory Design Training Set.

_5. Click on Build tab then lets retrain the Classification Module again.

_6. Click anywhere on Classification model.

_7. Click on Wage and Tax.

_8. Click Next button.

_9. Click Train button.

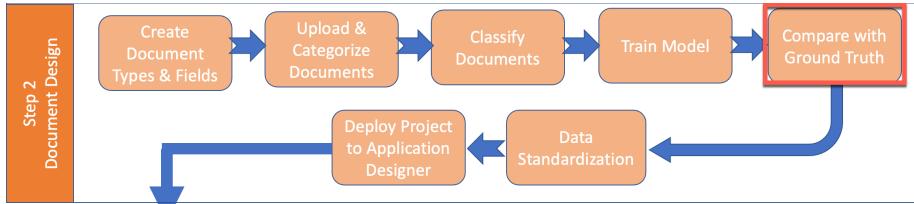
_10. Now look at the confidence score for Wage and Tax.

_11. Click Next and then Click Done

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.

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.

- 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.



- 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.

Automation Document Processing Lab

The screenshot shows the 'Extraction model' step in the 'Business automation / User01-CEB /' workflow. The 'Training set (3)' contains three documents: TR_FW2_1001_0000_P5.pdf, TR_FW2_2000_0000_P5.pdf, and TR_FW2_3001_0000_P5.pdf. The 'Test set (2)' contains two documents: TR_FW2_3000_0000_P5.pdf and TR_FW2_4000_0000_P5.pdf. A message box at the top right says: 'Please make sure you have at least 1 reviewed document to train the model. Review your training and test sets. A good practice is to assign 70% of your samples to the training set and 30% to the test set. The test set is used to generate the model training results.' Below the sets, there is an 'Auto generate 70/30 split' button.

_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**.

The screenshot shows the 'Teach model' step in the same workflow. The 'Teach Samples' button is highlighted with a red box. The table below lists three documents: TR_FW2_1001_0000_P5.pdf, TR_FW2_2000_0000_P5.pdf, and TR_FW2_3001_0000_P5.pdf, all marked as 'Not ready' with 0/7 fields reviewed.

Document name	Status	Fields reviewed	Date added
✓ TR_FW2_1001_0000_P5.pdf	Not ready	0/7	09/29/2022, 11:03 am
✓ TR_FW2_2000_0000_P5.pdf	Not ready	0/7	09/29/2022, 11:03 am
✓ TR_FW2_3001_0000_P5.pdf	Not ready	0/7	09/29/2022, 11:09 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.

Field Name	Value Captured
Federal Income Tax Withheld	123456789.99
State Income Tax Withheld	123456789.99
Social Security Wages	123456789.99
Medicare Wages	123456789.99
Alimony Wages	123456789.99

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.

Automation Document Processing Lab

- _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 a W-2 Wage and Tax Statement document. A pop-up window is open over the document, specifically targeting the 'Federal Income Tax Withheld' field. The field label 'Federal Income Tax Withheld' is underlined in blue, and a green checkmark is visible next to the captured value '123456789.99'. The pop-up also contains other fields for Social Security wages, Medicare wages, and tips, all of which have green checkmarks indicating they are complete.

- _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.

This screenshot shows the same W-2 form after the user has selected 'Save match' in the previous step. The 'Federal Income Tax Withheld' field now has a green checkmark next to it, and a blue message bubble indicates that the value has been saved ('Saved!'). The other fields in the pop-up window also have green checkmarks, signifying they are now complete.

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.

Automation Document Processing Lab

- _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. **Click** on the field label (“Employee’s first name and initial”). Again, **Click** on **Save match**

The screenshot shows the IBM Cloud Pak Administration interface. On the left, the W-2 Wage and Tax Statement form for 2020 is displayed. The form includes fields for employee identification number (22222), Social Security number (57-22-3048), and OMB No. (1441-0208). It also lists employer information: Long Lengthy Name The Corporation, 50334 Full Street Avenue Unit 1234, Minneapolis, Minnesota 55111-1234. The form is divided into sections for federal income tax withheld, state tax withheld, and local tax withheld.

On the right, a side panel titled "Field Name" and "Value Captured" is open. It shows a list of captured values for various fields. The "Employee Name and Address" section is expanded, showing the field label "e Employee's first name and initial" with a blue outline, indicating it is selected. Below it, the "Field value" section has a "Draw" button highlighted with a green box, and the text "e Employee's first name and initial" is shown.

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**

Automation Document Processing Lab

The screenshot shows a split-screen view. On the left, a W-2 form is displayed. On the right, a "Capture" interface is shown, allowing users to extract data from the document.

W-2 Form Data:

- Employee Social Security Number: 577-22-3048
- Employer Identification Number (EIN): 44-02328
- Employee's Name and ZIP code:
Long Lengthy Name The Corporation
56334 Full Sized Avenue Unit 1234
Minneapolis, Minnesota 55411-1234
- Driver number: 123456 A7B
- Employer's Address and ZIP code:
Michael Robert David Smithson III
56334 Full Sized Avenue Unit 1234
Minneapolis, Minnesota 55411-1234
- State: MN
- State wages, tips, etc.: 123456789.99
- State income tax: 123456789.99
- Local wages, tips, etc.: 123456789.99
- Local income tax: 123456789.99
- Local total: ABCDEFGH
- Year: 2020
- Department of the Treasury – Internal Revenue Service
- Copy 1 – For State, City, or Local Tax Department

Capture Interface:

- Field Name:** Value Captured
- Federal Income Tax Withheld:** 123456789.99
- Employee Name and Address:** Michael Robert David Smithson III
56334 Full Sized Avenue Unit 1234
Minneapolis, Minnesota 55411-1234
- Employee Social Security Number:** 577-22-3048
- Employer Identification Number:** 44-02328
- Mark this document as ready for training:**

- _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

Automation Document Processing Lab

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 like Employee Social Security Number (577-22-3048), Employer Identification Number (14-023285), and various income and tax amounts. On the right, a table lists captured fields with their values. A red arrow points to the 'Mark this document as ready for training' checkbox at the bottom of the right pane.

Field Name	Value Captured
Federal Income Tax Withheld Required	123456789.99
Employee Name and Address Required	
Employee Social Security Number	577-22-3048
Employer Identification Number	14-023285
Employers Name and Address	Long Lengthy Name The Corporation 56334 Full Street Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
Social Security Wages	123456789.99
Wages Tips Other Compensation	123456789.99

_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**

Automation Document Processing Lab

The screenshot shows the IBM Cloud Pak Administration interface. On the left, there are two W-2 forms for the year 2020. The top form is for Michael Robert David Smithson III and the bottom form is for AAA BBB CCC. Both forms show various tax withholdings and employer information. To the right of the forms is a data extraction tool. It has a table titled "Field Name" and "Value Captured" with several entries. Below the table, there are input fields for "Field label (optional)" containing "1 Wages, tips,othercomp." and "Field value" containing "123456789.99". At the bottom of the interface, there is a button labeled "Mark this document as ready for training."

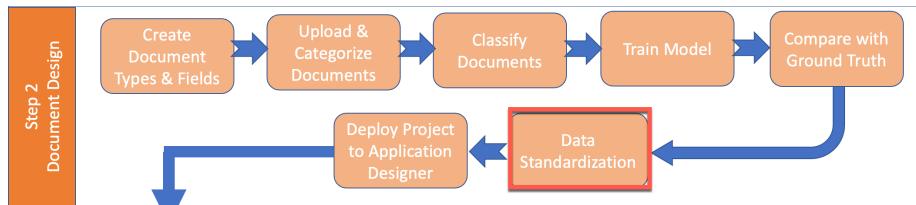
8.2 Train extraction model

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

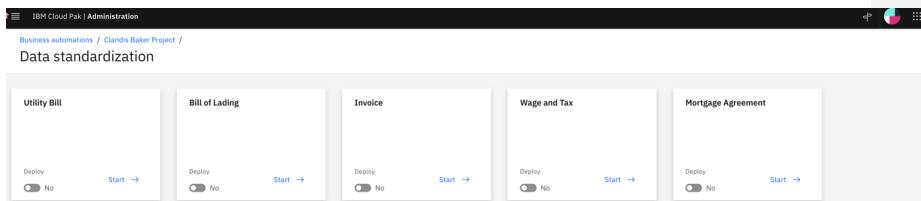


Next, we may need to standardize the data that will be presented in the user interface and how it will be stored in the FileNet repository for example. Data standardization is the process of defining attributes for a data field in a standardized way. This is done using data definitions. These definitions 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.

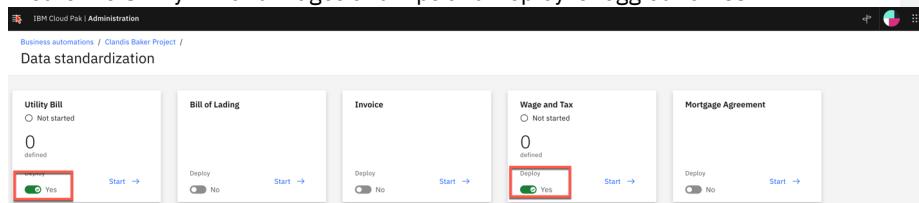
1. Return to the guided configuration flow and **Click** anywhere in the **Data standardization** box

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.

Automation Document Processing Lab



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



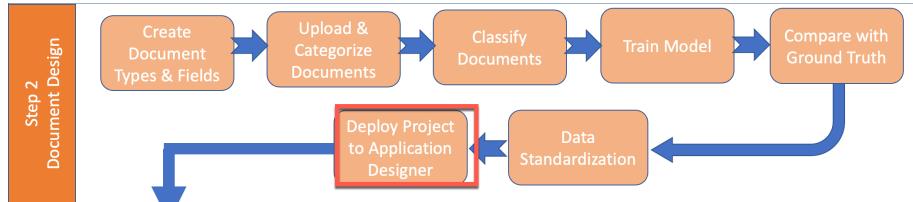
_3. Click on Start on either selected deployment.

This is where we begin defining the data field attribute definitions. You could create a new data definition and configure them. We will NOT be creating/defining any data fields for this lab.

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

[Business automations / Clandis Baker Project /](#)

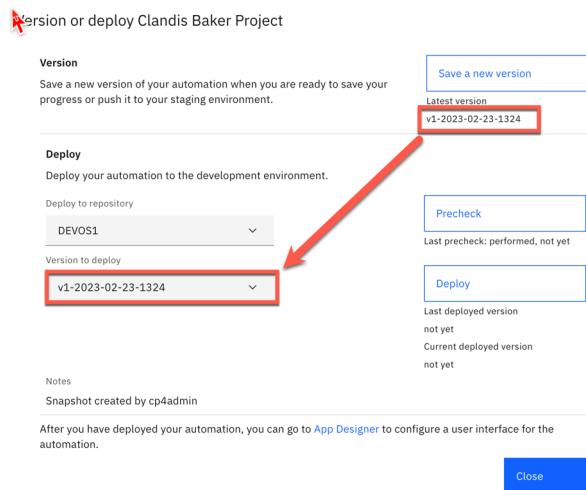
10 Version and deploy your project



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

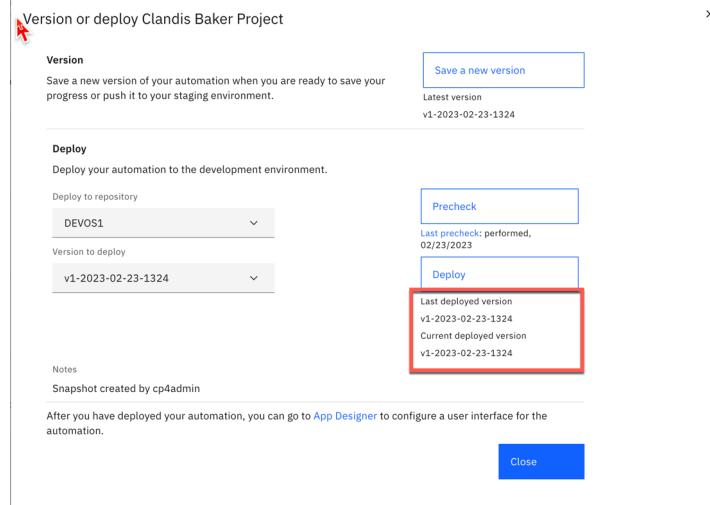


... also, in the top corner has the “Latest Version”

Automation Document Processing Lab

- _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.

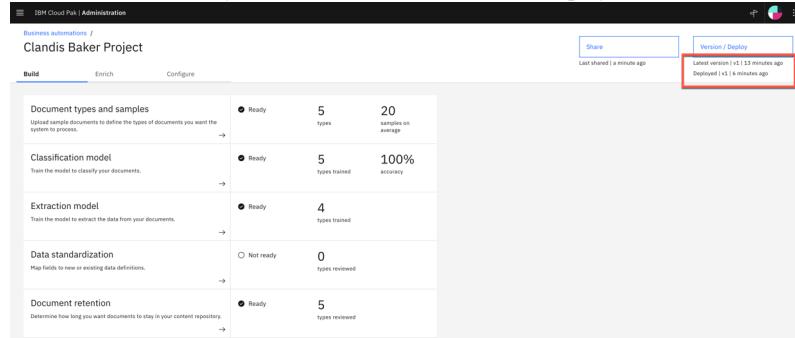


Note that you do not have to remain in the deploy screen while it is versioning or deploying. You can always click the button and then go back into any other screen if you like. It will run in the background. If you do this, just keep an eye on the top right of your screen for deployment status.

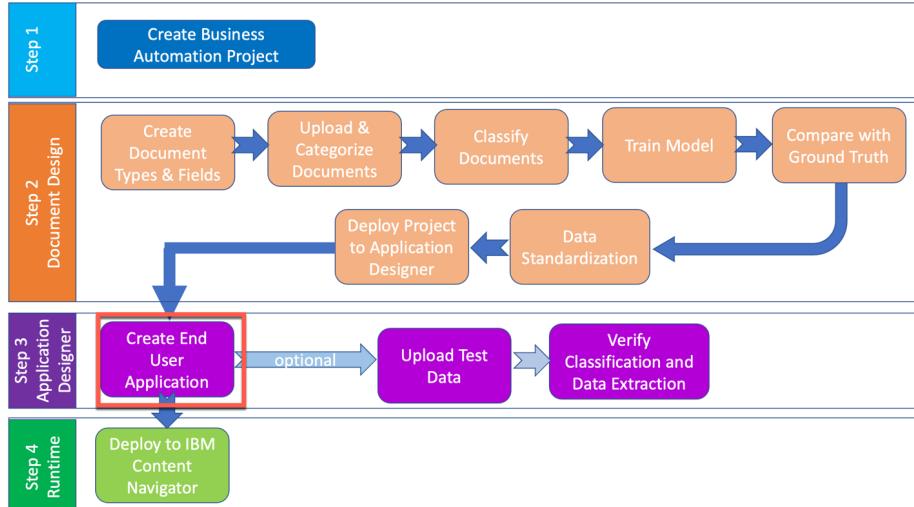
- _5. Click **Close** button.

Once deployed, proceed to the next steps.

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



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.

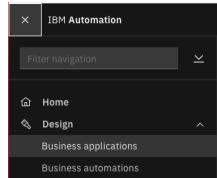
11.1 Create your Runtime Application.

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



and **selecting Business Applications**

Automation Document Processing Lab



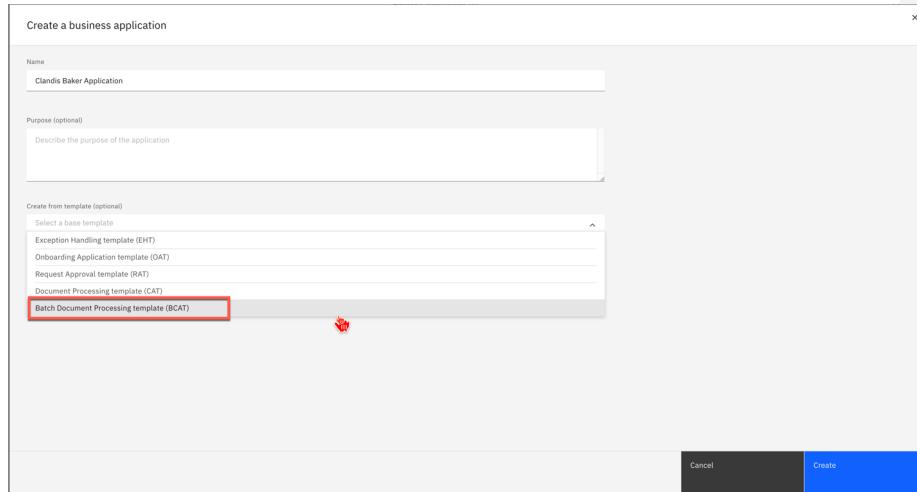
_2. From the **Create** drop down list, select Application

A screenshot of the IBM Cloud Pak Administration interface under the Business applications section. It shows a 'Create' dropdown menu with options: Application (selected), Template, Toolkit, and Toolkits. Below the menu, there's a message: 'There are no applications to display...yet.' with a link to 'Create'. Three template cards are displayed: Request Approval template, Onboarding Application template, and Exception Handling template. Each card includes a description, last update date (02/20/2023), and a red 'Create' button.

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

A screenshot of the 'Create a business application' dialog box. The 'Name' field is filled with 'Clandis Baker Application' (highlighted with a red box). The 'Purpose (optional)' field contains the placeholder 'Describe the purpose of the application'. Under 'Create from template (optional)', there's a dropdown menu with several options: 'Exception Handling template (EHT)', 'Onboarding Application template (OAT)', 'Request Approval template (RAT)', 'Document Processing template (CAT)', and 'Batch Document Processing template (BCAT)'. At the bottom right of the dialog are 'Cancel' and 'Create' buttons, with 'Create' being highlighted in blue.

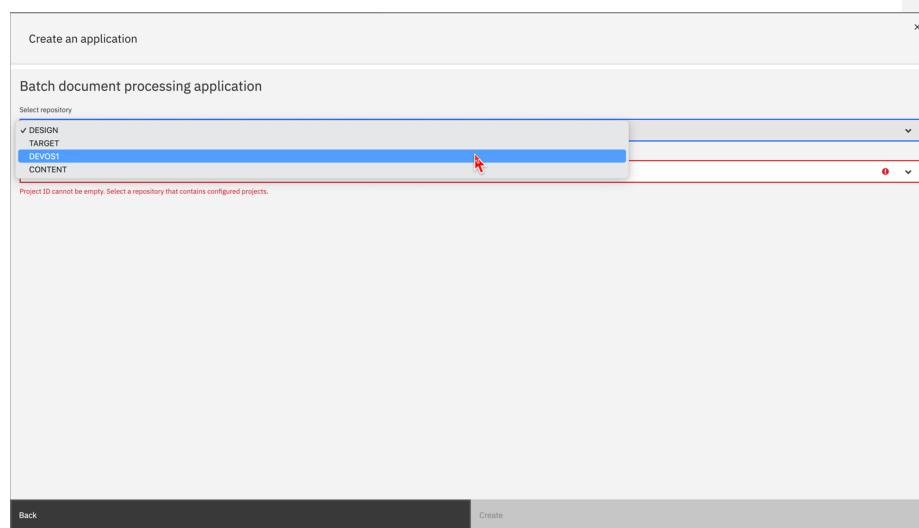
- _4. In the Create Form Template give it a <Name> and in drop down **select Batch Document Processing template (BCAT)**.



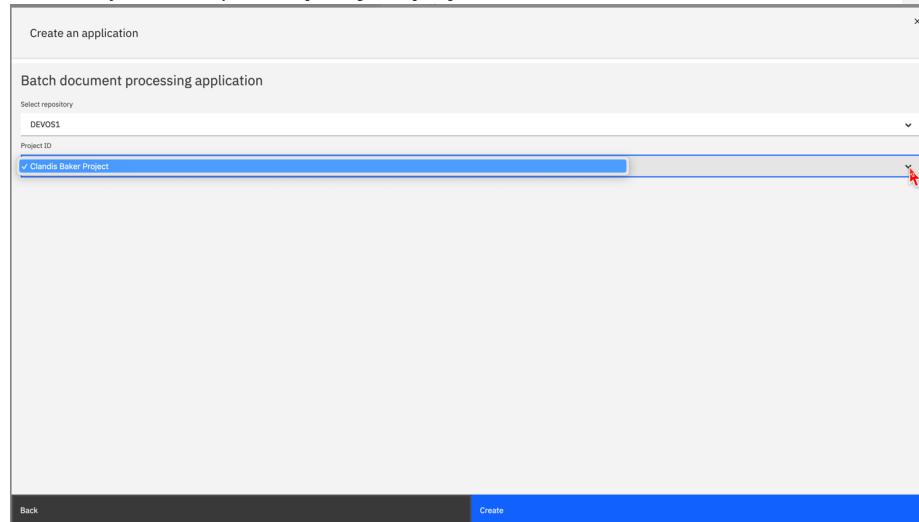
You could have selected the Document Processing Template if you only wanted to process a single document at a time, but in this lab, you will process several documents in a batch.

- _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**

Automation Document Processing Lab

You should now be in the *Application Designer*

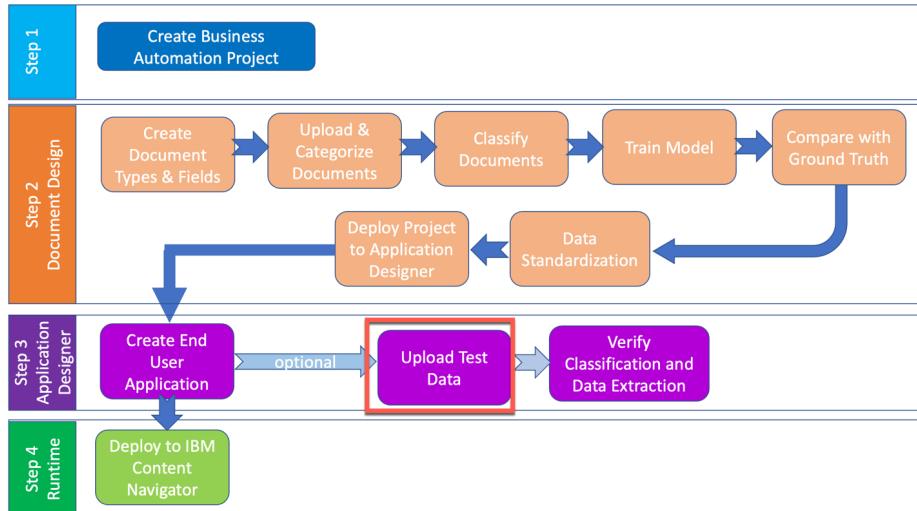
The screenshot shows the IBM Cloud Pak Application Designer interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Administration' and 'Business applications / Cländis Baker Application'. A red arrow points to the application name. Below the navigation is a toolbar with 'Content' (selected), 'Grid', and other icons. To the right is a 'Preview' button. The main area has a header 'Last saved seconds ago by you.' and a 'Drag a component to your page' panel on the right containing various UI components like 'Add batch modal', 'Add document modal', 'Add folder modal', 'Batch content', etc. On the left, there are sections for 'Review batch issues' (Document type and page order issues, Data extraction issues) and 'Batches' (Content List table). The table shows three documents: My Document1, My Document2, and My Document3, with columns for Name, Size, Modified by, Last modified, and Version. The table includes standard CRUD buttons ('Add', 'Edit', 'Delete') and pagination controls ('Items per page: 100', 'Items 1-3').

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.

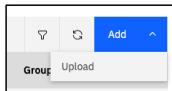
The screenshot shows the "Review batch issues" section of the application. It displays two categories of issues:

- Document type and page order issues: 0 batches
- Data extraction issues: 0 batches

Below this, there is a "Batches" table with the following columns: Name, Files, Priority, Status, Added on, Added by, Group, and Location. A search bar and a message stating "No items found." are also visible.

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.

A screenshot of a 'Upload new batch' form. It includes fields for 'Display Name' (Batch 1), 'Description', and 'Priority' (High). The 'Priority' field has a dropdown menu open.

- _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

Automation Document Processing Lab

_6. Click on the Add button.

The screenshot shows the 'Review batch issues' section with two tiles: 'Document type and page order issues' (0 batches) and 'Data extraction issues' (0 batches). Below this is a table titled 'Batches' with columns: Name, Files, Priority, Status, Added on, Added by, Group, and Location. A single row is visible for 'Batch01'. At the bottom left is a pagination control 'Items per page: 100 1-1 of 1 items'. The top right has a link 'Learn more about document processing'.

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

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

The screenshot is similar to the previous one, showing the 'Review batch issues' section and the 'Batches' table. The 'Status' column for 'Batch01' now shows 'Documents uploaded'. The top right still has the 'Learn more about document processing' link.

_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.

The screenshot shows the 'Review batch issues' section with the same two tiles. The 'Batches' table now shows a single row for 'Batch 1'. The status column for 'Batch 1' shows 'Document issues' with a timestamp '01/13/2021, 08:44 am' and 'CEAdmin' as the 'Added by' user. The top right has the 'Learn more about document processing' link.

11.3 Correct any classification errors.

_1. Click on the Document type and page order issues tile to open the batch.

Automation Document Processing Lab

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

Commented [CB2]: May need new screen

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	Utility Bill

Issues (1 of 5)

Review document type

Dismiss

1

20222 577-22-0006 Date to: 0000

1 Wage and Tax statement 1000.00

1 Miscellaneous income 1113.33

1 Wages and salaries 1770.00

1 Miscellaneous income 241.86

1 Total taxable 4000.00

1 Miscellaneous income 443.21

1 Miscellaneous income 256.00

1 Miscellaneous income 2000.00

1 Total taxable 332.00

1 Miscellaneous income 420.86

1 Department's address and DIF code 1000.00

1 State taxes 200.00

1 City taxes 200.00

1 Total taxable 1770.00

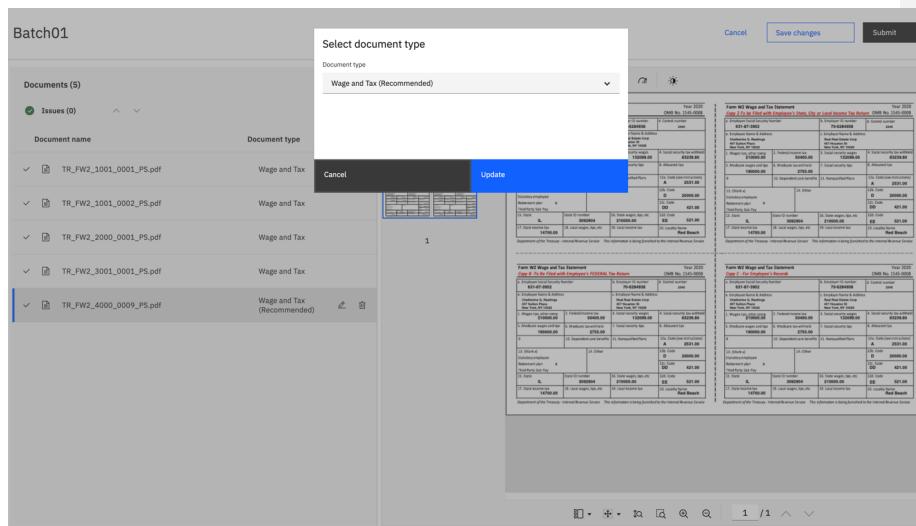
W-2 Wage and Tax Statement 2020 Department of the Treasury - Internal Revenue Service

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

Automation Document Processing Lab



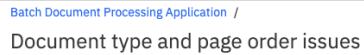
- _5. Review all documents to ensure everything is correct. If the system no longer detects any issues, you should see a green checkmark near the top of the document list.



- _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.



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.

A screenshot of a user interface showing a single item in a list. The item is labeled "Data extraction issues" with a small info icon to its right. Below it, it says "1 batches".

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

A screenshot of a user interface showing a list of items under "Batch 1". The first item is "Batch 1".

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

A screenshot of a user interface showing a table of documents. The columns are "Name", "Issues", "Status", "Modified on", and "Modified by". There are five rows, each representing a document. The "Issues" column shows yellow triangles indicating data issues. The "Status" column shows "Data issues" for most documents, except for the last one which shows "Issues reviewed". The "Modified on" and "Modified by" columns show the date as 23/02/2023 and the user as cp4admin.

Name	Issues	Status	Modified on	Modified by
TR_FW2_1001_0001_P5.pdf	1	Data issues	23/02/2023	cp4admin
TR_FW2_1001_0002_P5.pdf	1	Data issues	23/02/2023	cp4admin
TR_FW2_2000_0001_P5.pdf	1	Data issues	23/02/2023	cp4admin
TR_FW2_3001_0001_P5.pdf	2	Data issues	23/02/2023	cp4admin
TR_FW2_4000_0009_P5.pdf	0	Issues reviewed	23/02/2023	cp4admin

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.

Automation Document Processing Lab

The screenshot shows a document processing application window. On the left is a preview of the W-2 tax form. On the right is a panel titled "Extracted data" showing various fields and their values. The fields include:

- Federal Income Tax Withheld:** 1800.00
- 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:** Test and Rest Inc., 563 Stoney Brook Rd Minneapolis, MN 55411
- Organization:** (none)
- Name:** (none)

Take a moment to discover the image viewer features.

Image viewer features at top:

The screenshot shows the same document processing application window. A red box highlights the toolbar icons at the top of the image viewer, which include options for rotation, zoom, and other visual effects.

- Rotate image
- Visual effect adjustment
- Invert

Automation Document Processing Lab

Image viewer features at bottom:

The screenshot shows a document processing application interface. On the left is a thumbnail view of the W-2 form. The main area displays the extracted data from the form, including fields like Employee Social Security Number (577-22-3048), Employer Identification Number (14-023285), and various tax amounts. A red box highlights the 'Fields with issues' section, which lists validation errors for the Employee Social Security Number and Employer Identification Number. Navigation controls at the bottom allow for page and thumbnail switching.

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

Field features

This screenshot shows the same document processing application interface as the previous one. It displays the W-2 form and its extracted data. A red box highlights the 'Fields with issues' section, which lists validation errors for the Employee Social Security Number and Employer Identification Number. The navigation controls at the bottom are identical to the first screenshot.

- 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 comparison between a scanned W-2 form and its digital extracted data representation. The W-2 form is on the left, and the extracted data is on the right. The extracted data includes various fields such as Employee Social Security Number (577-22-3048), Employer Identification Number (14-023285), and various monetary amounts. A red box highlights the 'Fields with issues' section, which lists two validation errors: 'Validation error' for the Employee Social Security Number and 'Validation error' for the Employer Identification Number. Both of these fields are marked with an asterisk (*) indicating they are mandatory.

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

This screenshot shows the 'Extracted data' view with a dropdown menu open. The menu has two options: 'All Fields' (which is highlighted with a blue background) and 'Fields with issues'. A red arrow points to the 'All Fields' option, indicating where the user should click.

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

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

Automation Document Processing Lab

TR_FW2_1001_0001_PS.pdf | ▲ Document type: Wage and Tax

Extracted data

All Fields

Federal Income Tax Withheld *

1800.00

Employers Name and Address

Test and Rest Inc. 543 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

Extracted data

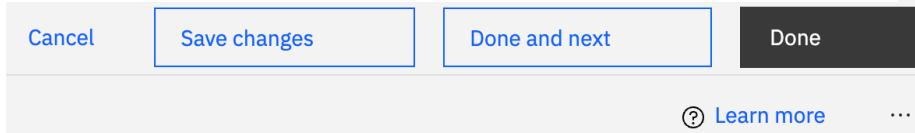
Fields with issues

Issue types

There aren't any extraction issues

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

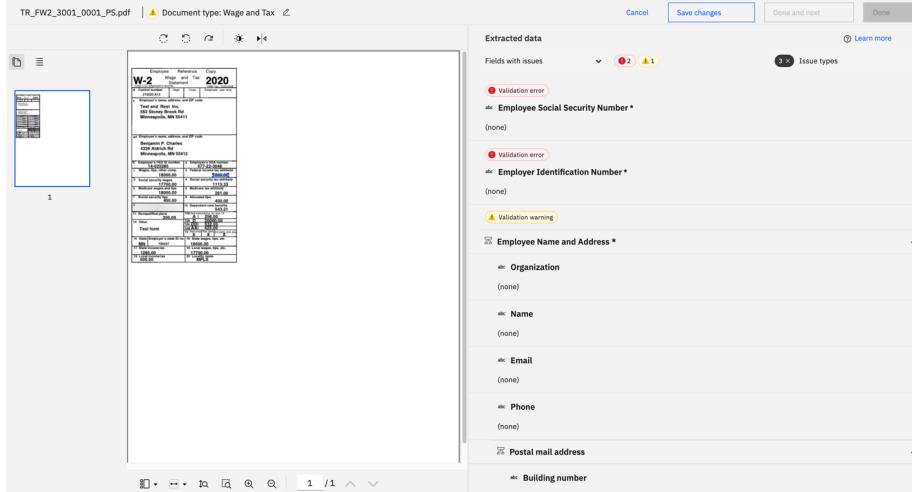
_6. Click on **Done and next** box at the top.



_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).



_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.

Automation Document Processing Lab

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

The screenshot shows a document processing application with a PDF viewer on the left and a data extraction interface on the right. The PDF contains a W-2 tax form for Benjamin P. Charles. The data extraction interface highlights the Social Security Number (SSN) field with a red box. The extracted data table includes fields like Employee's name, address, and ZIP code; Employer's FED ID number; and various tax withholdings and benefits. A validation error message 'Required value is missing.' is displayed next to the SSN field.

_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' screen in the IBM Cloud Pak Administration interface. At the top, there's a header bar with the title 'Document processing automation (1)'. Below the header, there's a section titled 'Business automations' with a brief description. A single project named 'Clandis Baker Project' is listed. Under the heading 'Published automation services', there are four categories: 'Decision', 'Document processing' (which is highlighted with a blue border), 'Workflow', and 'External'. At the bottom of the screen, there are 'Create' and 'Import' buttons.

_2. Select <your project name> Click open

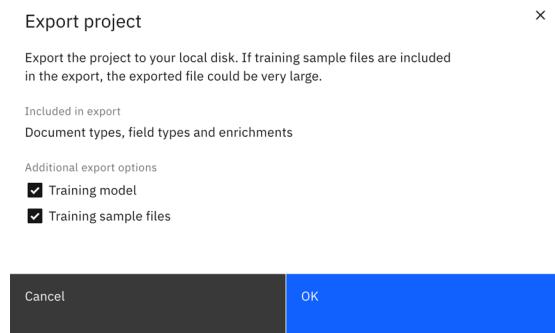
This screenshot is similar to the previous one, but the 'Clandis Baker Project' is now highlighted with a red box. The 'Document processing' category under 'Decision' is also highlighted with a blue border. The rest of the interface elements are identical to the first screenshot.

3. From the Main screen select the Configure tab

This screenshot shows the main screen for the 'Clandis Baker Project'. The top navigation bar has 'Business automations / Clandis Baker Project'. Below it, there are three tabs: 'Build', 'Enrich', and 'Configure', with 'Configure' being the active tab. On the right side, there are two buttons: 'Share' and 'Version / Deploy'. Below the tabs, there are sections for 'Import / Export ontology' (with 'Language settings' and 'Git server configuration' options), 'Export project' (with an 'Export project' button), and 'Import project' (with an 'Import project' button). At the very bottom, there are status indicators: 'Last shared 12 hours ago', 'Latest version v2 12 hours ago', and 'Deployed v2 12 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.

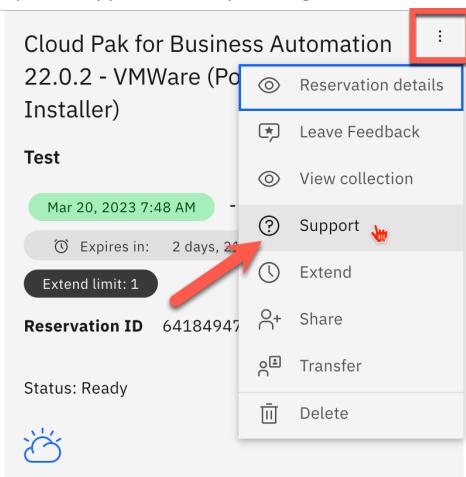
An operator fails to install and continuously shows Pending status.

For fix visit below link.

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

Other issue could be the deployment itself had an issue. Two things to do in this case.

1. Open a support ticket by clicking on the 3 dots on the tile.



IBM Internal can also access support via SLAC Channel at #itz-techzone-support

2. Delete tile and try to deploy again.

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.

Automation Document Processing Lab

```
#####
# 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) FwthMenu
#? 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

Console Automation Document Processing
#####

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
```

*****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:

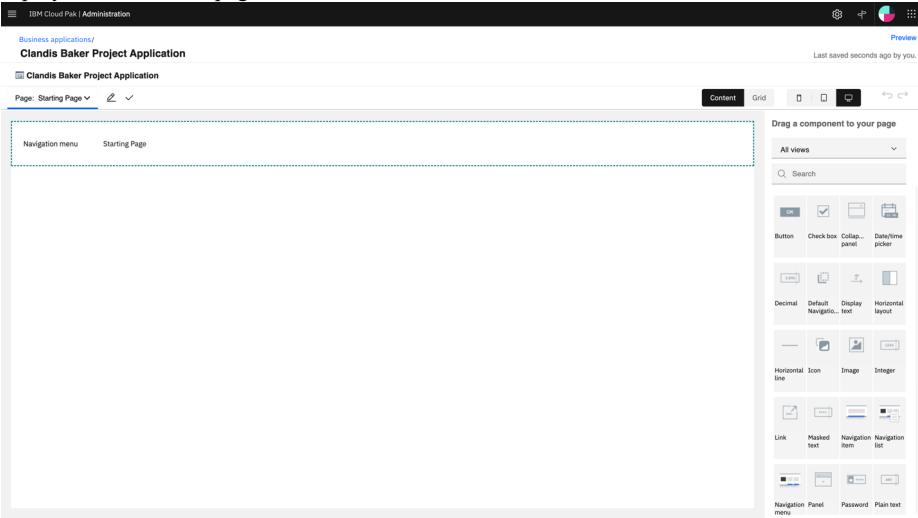
- Open the **OpenShift web console** in a browser.
- Click on **Workloads -> ConfigMaps** on the left-hand side navigator.
- Type ***access-info*** in the field next to 'Name'.

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

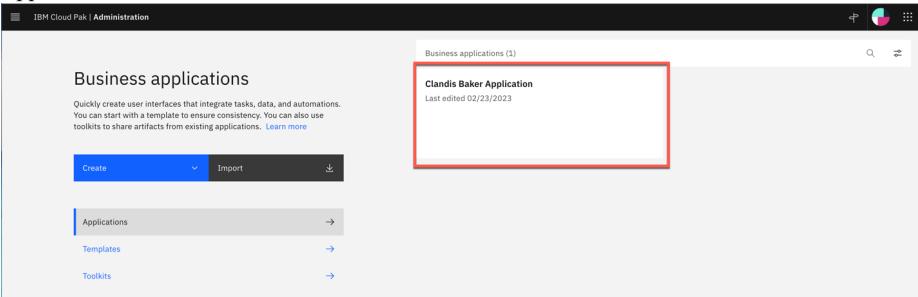
Automation Document Processing Lab

APPLICATION BLANK

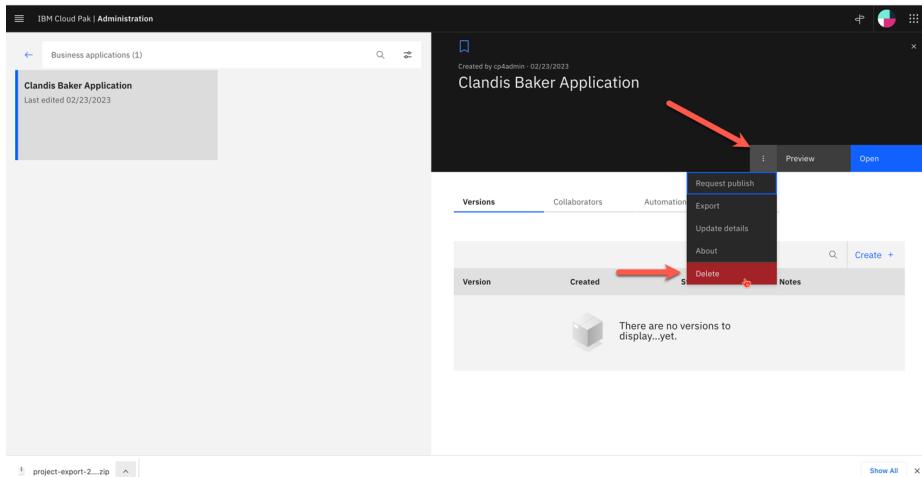
During creating of Business Application setup, sometimes on first time after project has been deployed. The Starter page is blank.



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.

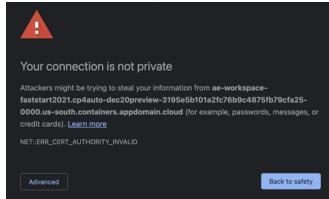
OPENING AN INCOGNITO WINDOW

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:

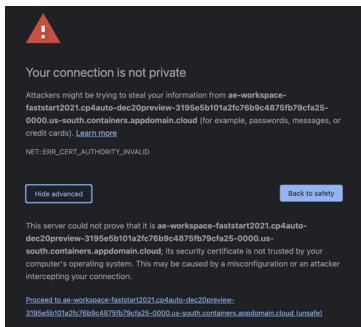
Open the Generate Security Tokens Box note and click all 3 of the links listed. This will reset the self-signed security certificates.

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

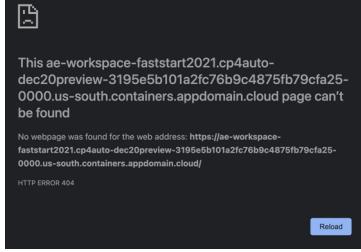


Click Advanced, and the browser window will look something like this:

Automation Document Processing Lab



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>

Appendix C - Badge Information.

Badge quiz page - <https://learn.ibm.com/course/view.php?id=12413>

Credly page - <https://www.credly.com/org/ibm/badge/ibm-automation-document-processing-tech-jam>