

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 Icons	6
1.3 Abstract	6
1.4 Introduction.....	7
2 Getting started	8
2.1 IBM TechZone – Overview	8
2.1.1 Reserve Environment	9
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.....	56
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.....	73
11.4 Correct extraction issues.....	75
12 Export Import Project	80
Appendix A - Troubleshooting	82
TechZone Pending Status taking Long Time	82
Can't find user/password in Daffy	82
APPLICATION BLANK	85
Connection issue with Workstation to Cloud.....	86

OPENING AN INCOGNITO WINDOW.....	86
Appendix B - BAW & ADP Integration Sample	88
https://github.com/IBM/baw-adp-integration-sample	88
Appendix C - Badge Information.	89

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

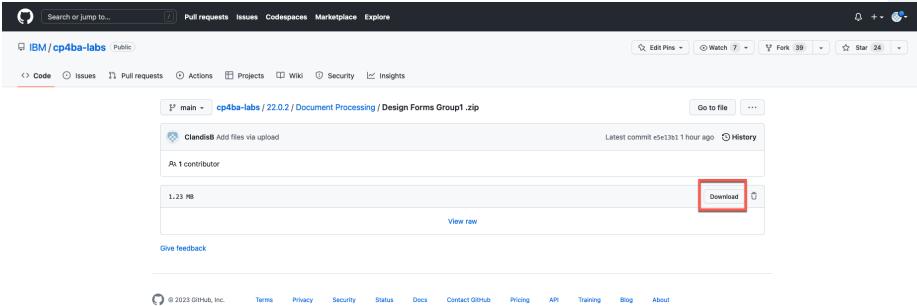
Download the sample documents in the zip file. We will be using these sample documents during the labs You can find them here:

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

Group 1 - Design Docs.zip	Add files via upload	1 minute ago
Group 2 - Design Classification docs.zip	Add files via upload	1 minute ago
Group 3 - Application Runtime Set.zip	Add files via upload	1 minute ago
Readme.md	Update Readme.md	18 hours ago
[In Process]Lab Guide - Automation Document Proce...	Add files via upload	18 hours ago

_1. Click on “Group1 – Design Docs.zip”.

_2. Then Click on Download



_3. Repeat above steps “Group 2 – Design Classification docs.zip” and “Group 3 – Application Runtime 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.

Automation Document Processing Lab

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>



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
- _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

Automation Document Processing Lab

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

Preferred Geography (required)

Choose a preferred geography

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

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

VPN Access (required)

✓ Disable
Enable



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**

Starter Service (required)

✓ all
content
content-decisions
decisions
docprocessing
workflow
script

- _12. Click Submit

Cancel **Reset** **Submit**



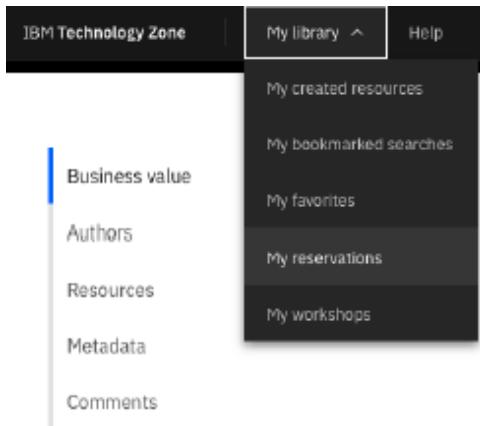
Upon receiving the Your environment is ready email, please allow up to 1 hour for the start-up services to fully complete. If after receiving email and a few hours have passed and your environment is not up, check [Appendix A – Trouble Shooting](#) for possible fix.

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>

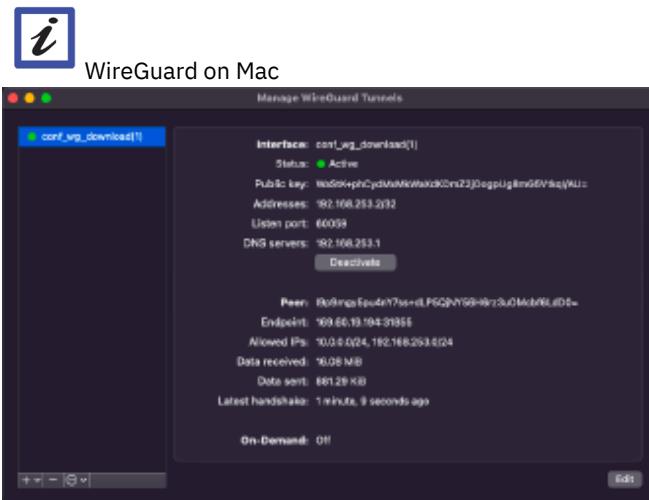
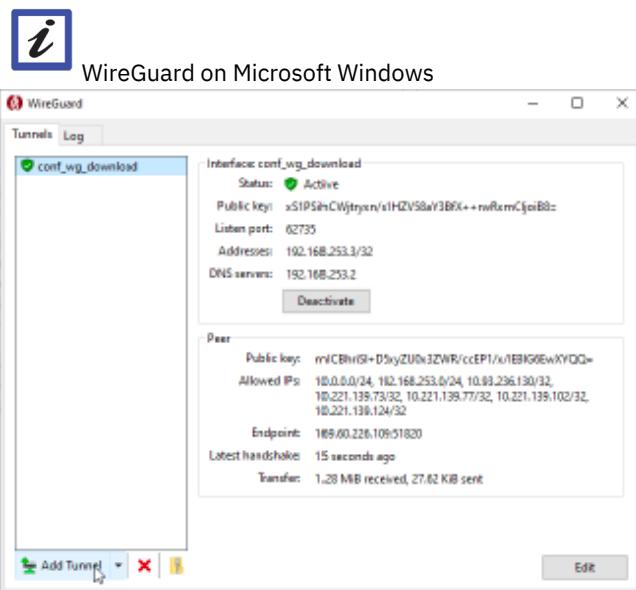
2.2 Set up WireGuard VPN

- _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](#)

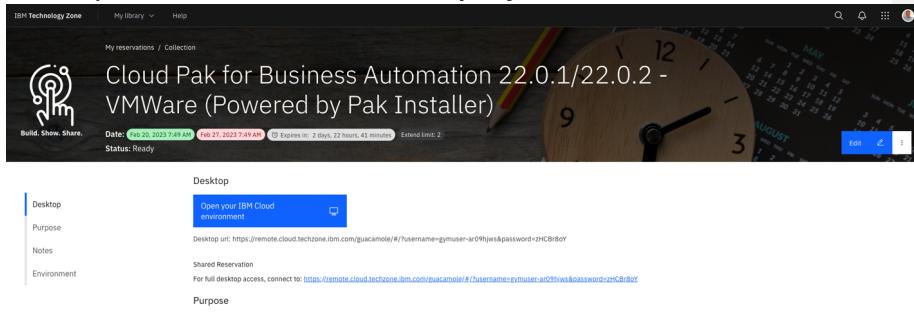
- _6. On your local workstation, install WireGuard by accessing <https://www.wireguard.com/install/>
- _7. Launch WireGuard
- _8. Click Add Tunnel and load the **conf_wg_download.conf** file.

Automation Document Processing Lab

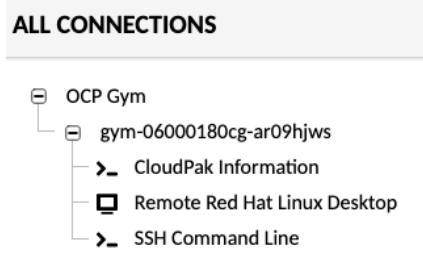


2.3 Open your IBM Cloud Environment

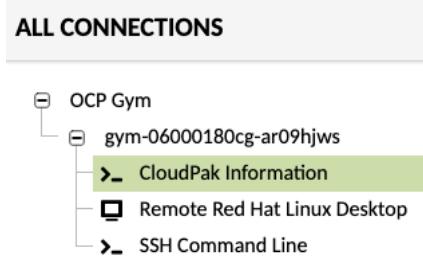
- _1. Back on your reservation screen **Click on Open your IBM Cloud environment**



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



- _3. **Select CloudPak Information**



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

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

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

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

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

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

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

Automation Document Processing Lab

```
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/ocpinstall/cp4ba/icp4adeploy-cp4ba-access-inf

Console Automation Document Processing
#####
Cloud Pak BusinessCard   : https://cp4ba-starter.apps.ocpinstall.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.

Automation Document Processing Lab

```
oc login https://api.ocpininstall.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
Zen Version : cp4ba-starter
Message 1 : 4.8.1
Message 2 : Running reconciliation
Message 3 : Prerequisites execution done.
Message 4 : FAIL - prerequisites Deployment failed ←
Deployment Service : Starter docprocessing
Config Map Dump : /data/daffy/log/ocpininstall/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.ocpininstall.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.

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.

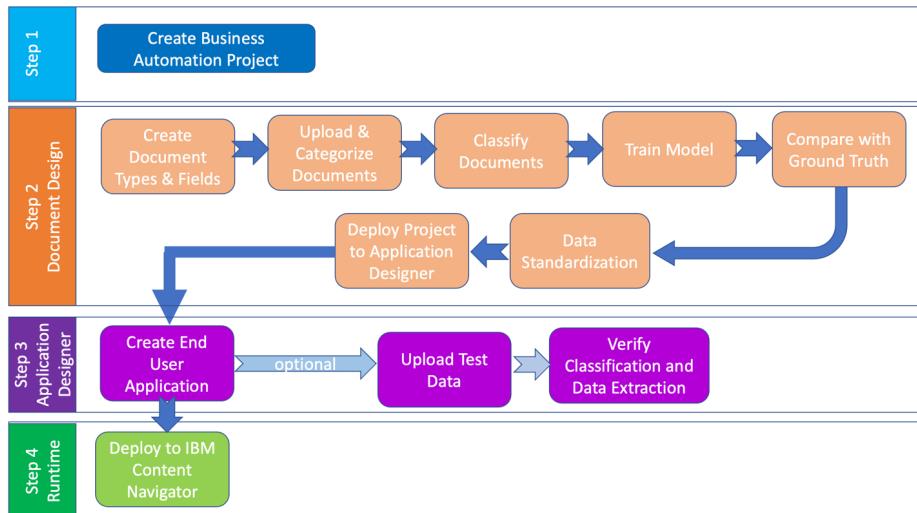
When you deploy your document processing application, you connect it to a content repository that manages the document types and the extracted data for each document. The solution is fully integrated with IBM FileNet® Content Manager, simplifying document and data storage by applying your existing filing 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.



Step 1 – Create Business Automation Project

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

Step 2 – Document Design

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

Step 3 – Application Designer

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

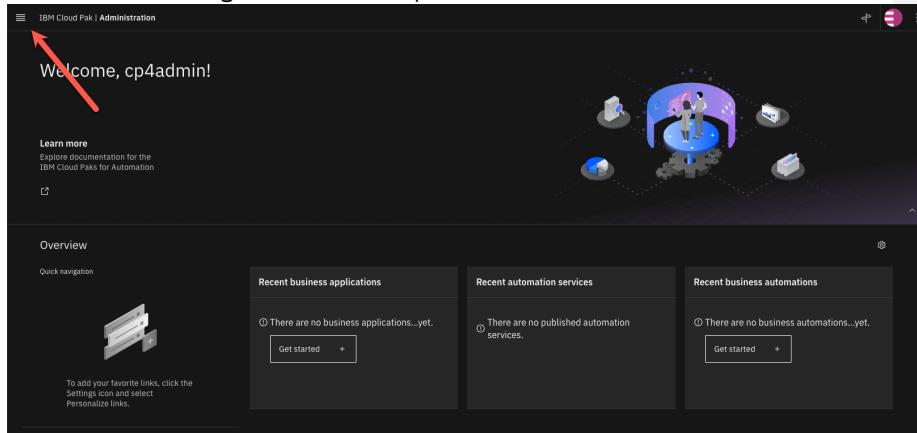
Step 1

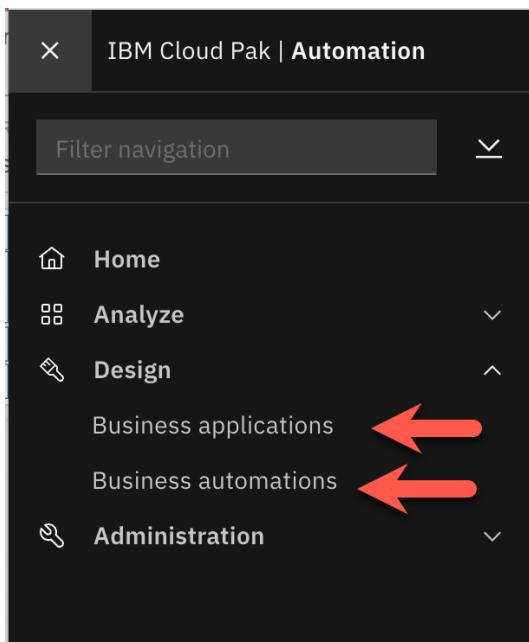
Create Business Automation Project

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.

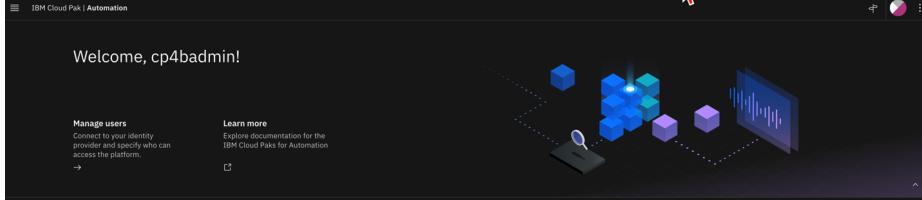
Within the Business Automations you can create or reuse automations. An automation is a collection of artifacts that fulfills a business purpose. You can publish some automation artifacts as automation services that you can call and reuse in a consistent way. Also in Business Automation, you use the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

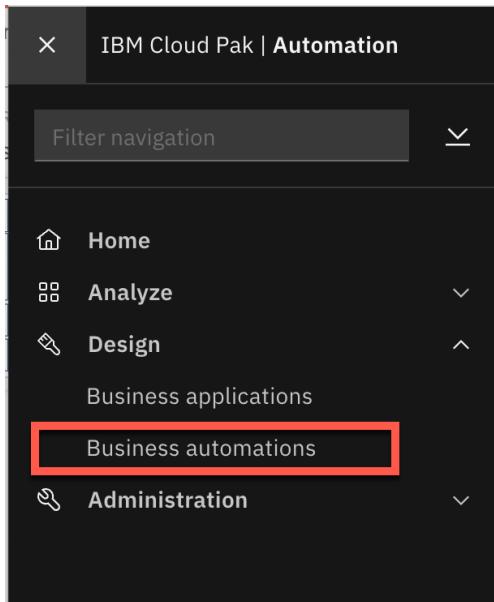
Within **Business Applications** you can quickly create user interfaces that integrate tasks, data, and automations. You can start with a template to ensure consistency. You can also use toolkits to share artifacts from existing applications.

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 the 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 has 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).

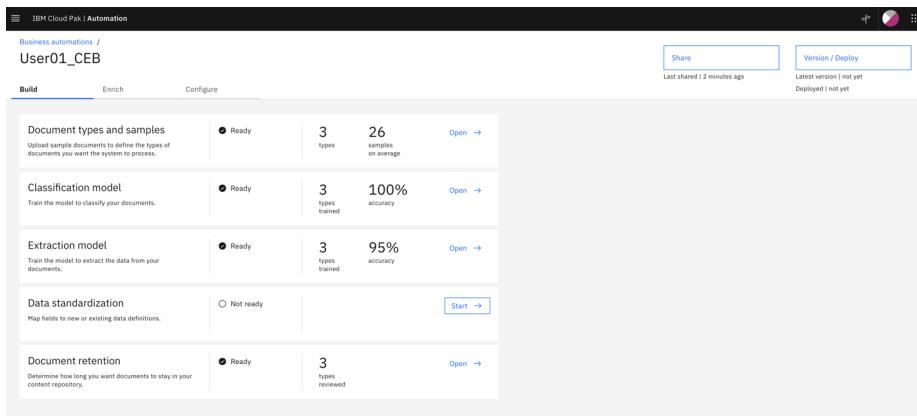
Automation Document Processing Lab

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.



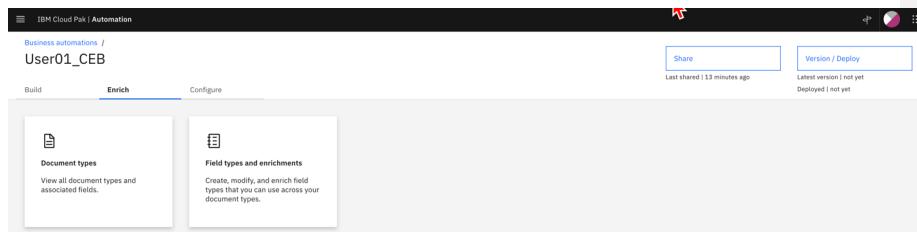
The screenshot shows the IBM Cloud Pak | Automation interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a search bar. Below it, a breadcrumb trail shows 'Business automations / User01_CEB'. On the right, there are 'Share' and 'Version / Deploy' buttons, with status information: 'Last shared 2 minutes ago', 'Latest version | not yet', and 'Deployed | not yet'. The main area has three tabs: 'Build', 'Enrich' (which is selected), and 'Configure'. The 'Enrich' tab displays several sections: 'Document types and samples' (3 types, 26 samples on average, 'Open'), 'Classification model' (3 types trained, 100% accuracy, 'Open'), 'Extraction model' (3 types trained, 95% accuracy, 'Open'), 'Data standardization' (Not ready, 'Start'), and 'Document retention' (3 types reviewed, 'Open').

4.1.2 Enrich Tab

_1. Click on the ENRICH tab.

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

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 'Cländis 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.

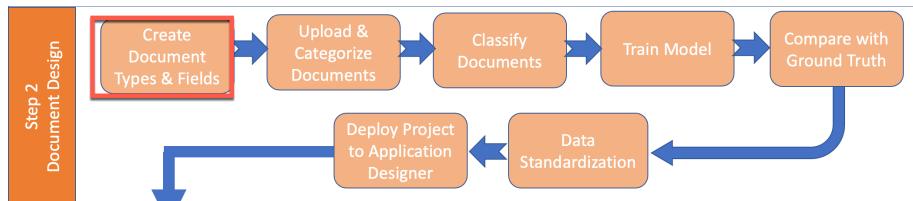
Automation Document Processing Lab

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.

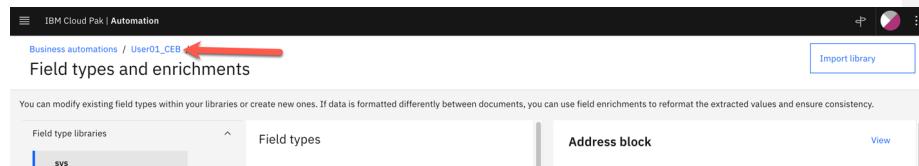


Before we use the guided configuration, you will configure some additional document types and fields used to extract data prior to uploading sample documents.

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.

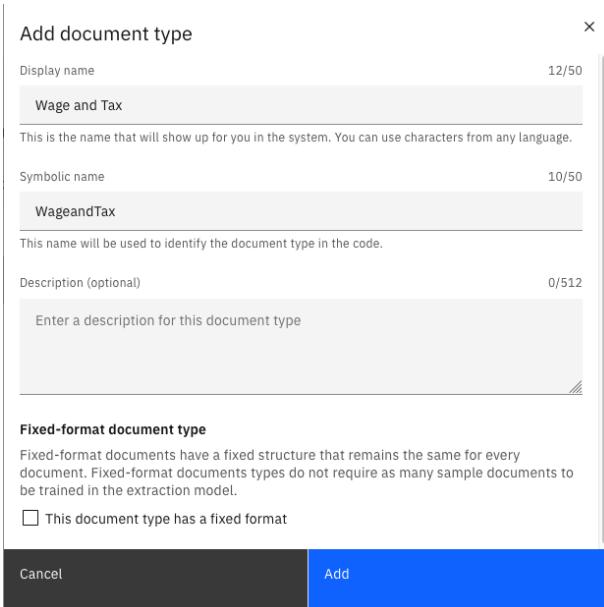
Symbolic name 10/50
WageandTax

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' listed. Under 'Document types', 'Bill of Lading', 'Invoice', and 'Utility Bill' are shown, while 'Wage and Tax' is selected. The main panel is titled 'Wage and Tax' and has tabs for 'General' and 'Attributes'. A message says, 'You haven't added any fields yet. Click the Add field button to get started.' Below this is a 'Add fields' button. The top right corner has a 'Create document type' 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

This screenshot is identical to the one above, showing the 'Wage and Tax' document type configuration. However, a red arrow points to the 'Add fields' button at the bottom of the central panel, highlighting the action to be taken.

_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:

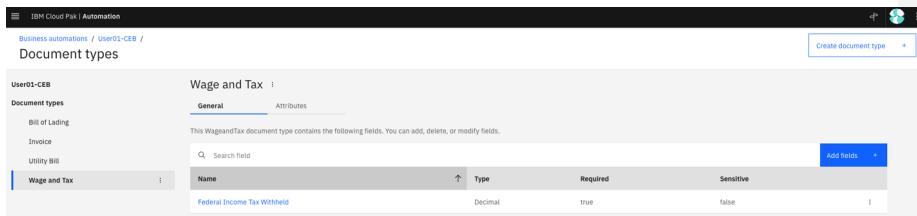
The screenshot shows the 'Create field' dialog in the IBM Cloud Pak | Automation interface. The document type selected is 'Wage and Tax'. The 'General' tab is active. In the 'Display name' field, the value 'Federal Income Tax Withheld' is entered. In the 'Aliases' section, two entries are present: '2 Federal income tax withheld' and '2. Federal income tax', 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

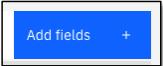
Automation Document Processing Lab

The screenshot shows the 'IBM Cloud Pak | Automation' interface. In the top navigation bar, the path 'Business automation / David-CustTest-Kiosk / Document Types / Employee name and address' is visible. Below the path, it says 'Document type: Form W2'. There are four tabs at the top: 'General' (selected), 'Field patterns', 'Value settings', and 'Subfields'. The 'General' tab has several sections: 'Display name' (Employee name and address, 25/10), 'Description (optional)' (Enter a description for this field, 0/512), 'Symbolic name' (Employee name and address), 'Aliases' (Enter an alternative name, 0/50), and 'Field type' (sys:Address information). Under 'Field type', there are two checkboxes: 'This field is required' (checked) and 'This field contains sensitive information' (unchecked). At the bottom of the General tab, there is a list of aliases: 'Employee name and address' (with a delete icon), 'Employee's first name and initial Last name Suffix' (with a delete icon), 'Employee's name, address, and ZIP code' (with a delete icon), 'Employee's name, address, and ZIP code' (with a delete icon), 'Employee Name & Address' (with a delete icon), and 'Employee's first name and initial' (with a delete icon).

- _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 at the bottom: 'This field is required' (checked) and 'This field contains sensitive information' (unchecked). The 'Value settings' tab is partially visible on the right.

_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

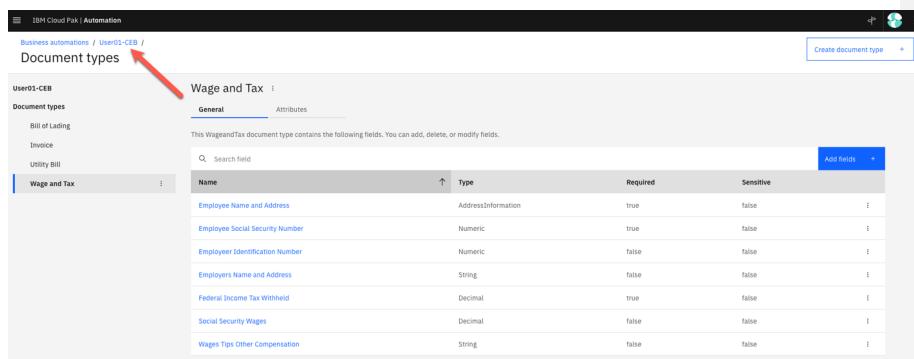
Reference for various field types:



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>

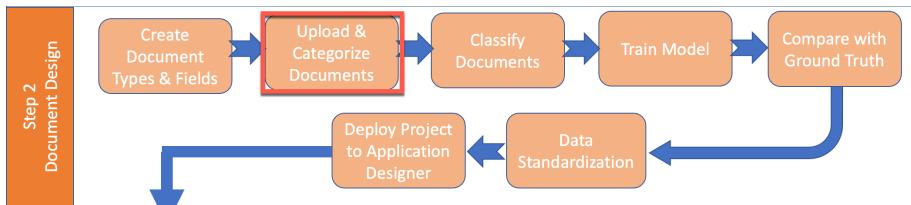
- _6. Click on the <name of your project> in the breadcrumb link in the top left of your screen. In the following example the name of the project is <User01_CEB>.



The screenshot shows the 'Document types' section of the IBM Cloud Pak for Automation interface. The breadcrumb navigation bar at the top displays 'IBM Cloud Pak | Automation' and 'Business automations / User01-CEB /'. Below this, the main title 'Document types' is followed by a sub-section title 'Wage and Tax'. There are two tabs: 'General' (selected) and 'Attributes'. A note below the title states: 'This WageandTax document type contains the following fields. You can add, delete, or modify fields.' A search bar and a 'Add fields' button are also present. On the left, a sidebar lists other document types: 'User01-CEB', 'Document types', 'Bill of Lading', 'Invoice', 'Utility Bill', and 'Wage and Tax' (which is currently selected). The main content area shows a table of fields:

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. To its right, it shows "4 types" and "22 samples on average". Below this, there are three more sections: "Classification model" (3 types trained, 100% accuracy), "Extraction model" (3 types trained, 97% accuracy), and "Data standardization" (0 types trained, Not ready).

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

Let's see what that looks like.

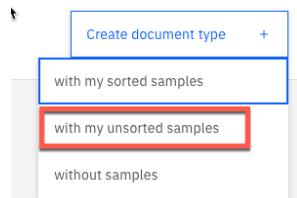
2. Click on **CREATE DOCUMENT TYPE** in the top right of the screen.

The screenshot shows a dropdown menu. At the top, it says "Create document type +". Below that, there are three options: "with my sorted samples" (which is highlighted with a blue box), "with my unsorted samples", and "without samples".

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

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

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



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

_4. Click Upload to get document samples.

From the downloaded sample documents open the folder name [Group 1 – Design Docs](#)

Note: this will take several 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 'Create document types'. At the top, there are tabs for 'Business automation / User interface / Document types and samples / Create document types'. Below the tabs are two buttons: 'Upload unsorted documents' and 'Review categories'. A note below the tabs says: '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 table lists 12 uploaded documents, each with a checkbox and a preview thumbnail. The documents are:

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_081_1_11.pdf
<input type="checkbox"/> UBILLCable_082_1_11.pdf

At the bottom of the table, there are buttons for '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.

Automation Document Processing Lab

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.

The image contains two screenshots of the IBM Cloud Pak Automation interface, specifically the 'Create document types' section. Both screenshots show a note about file naming and a list of documents categorized under 'Category 1 sample documents'.

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.*

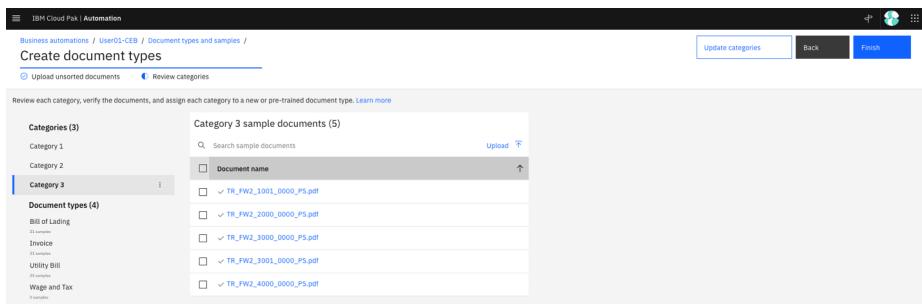
Screenshot 1 (Top): Category 1 sample documents (2). Contains two PDF files: UBILLCable_081_1_1.pdf and UBILLCable_082_1_1.pdf.

Document name
UBILLCable_081_1_1.pdf
UBILLCable_082_1_1.pdf

Screenshot 2 (Bottom): 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.

Document name
Mortgage Agreement1.pdf
Mortgage Agreement2.pdf
Mortgage Agreement3.pdf
Mortgage Agreement4.pdf
Mortgage Agreement5.pdf

Automation Document Processing Lab

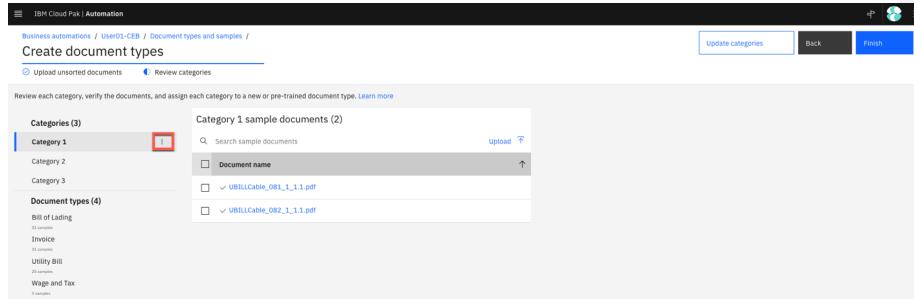


The screenshot shows the 'Create document types' interface in IBM Cloud Pak | Automation. The 'Categories' section lists three categories: Category 1, Category 2, and Category 3. The 'Document types' section lists four types: Bill of Lading, Invoice, Utility Bill, and Wage and Tax. Under Category 3, there is a sub-section titled 'Category 3 sample documents (5)' which 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.



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

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



The screenshot shows the 'Create document types' interface in IBM Cloud Pak | Automation. The 'Categories' section lists three categories: Category 1, Category 2, and Category 3. The 'Document types' section lists four types: Bill of Lading, Invoice, Utility Bill, and Wage and Tax. Under Category 1, there is a sub-section titled 'Category 1 sample documents (2)' which contains two PDF files: UBILLCable_081_1_1.pdf and UBILLCable_082_1_1.pdf. A red box highlights the three-dot menu icon next to 'Category 1'.

- _8. Select ASSIGN TO DOCUMENT TYPE**

Automation Document Processing Lab

The screenshot shows the 'Create document types' page in the IBM Cloud Pak | Automation interface. On the left, there's a sidebar with 'Categories (3)' and 'Document types (4)'. Under 'Categories', 'Category 1' has a red box around its 'Assign to document...' button. Under 'Document types', 'Bill of Lading', 'Invoice', 'Utility Bill', and 'Wage and Tax' are listed. On the right, there's a search bar and a list of 'Category 1 sample documents (2)' with checkboxes.

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

The screenshot shows the 'Assign documents' dialog box. It has a radio button for 'New document type' and 'Existing document type' (which is selected). Below is a dropdown menu for 'Utility Bill' containing 'Bill of Lading', 'Invoice', 'Utility Bill' (selected), and 'Wage and Tax'. At the bottom are 'Cancel' and 'Assign' buttons.

- _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

Assign documents

Assign documents of Category 2 to

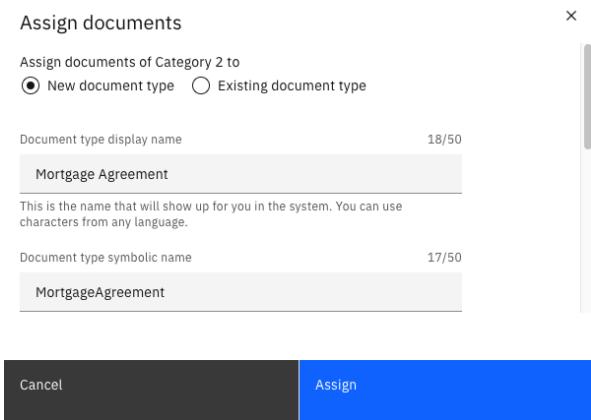
New document type Existing document type

Document type display name 18/50
Mortgage Agreement

This is the name that will show up for you in the system. You can use characters from any language.

Document type symbolic name 17/50
MortgageAgreement

Cancel **Assign**



- _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.

Assign documents

Assign documents of Category 3 to

New document type Existing document type

Document types

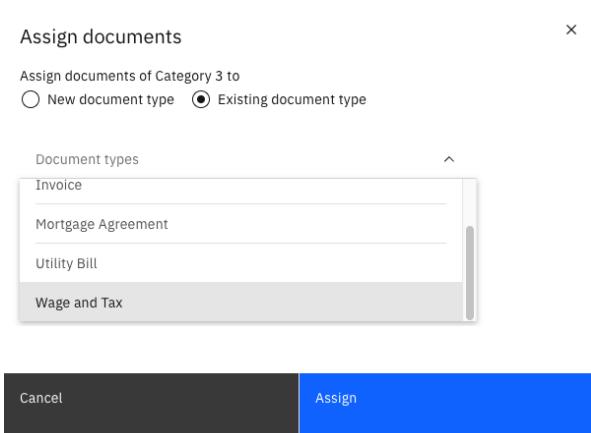
Invoice

Mortgage Agreement

Utility Bill

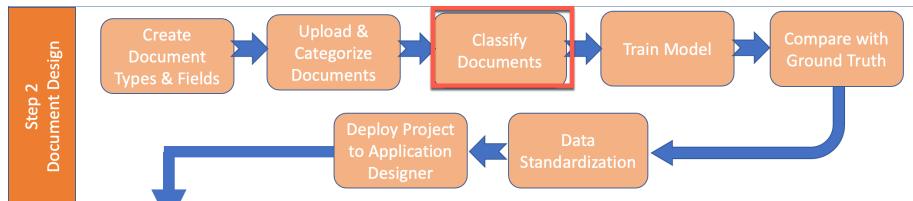
Wage and Tax

Cancel **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%
Last trained 4 days ago

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

Document types

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

Mortgage Agreement

5 samples

Mortgage Agreement sample documents (5)

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

Training set (5) 100% of total samples

Test set (0) 0% of total samples

Auto generate 70/30 split

Search training set sample documents

Mortgage Agreement1.pdf
Mortgage Agreement2.pdf
Mortgage Agreement3.pdf
Mortgage Agreement4.pdf
Mortgage Agreement5.pdf

Search test set sample documents

Mortgage Agreement1.pdf
Mortgage Agreement2.pdf

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

Classification model Accuracy: 84.8%
Last trained 4 days ago

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

Document types

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

Mortgage Agreement

5 samples

Mortgage Agreement sample documents (5)

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

Training set (3) 60% of total samples

Test set (2) 40% of total samples

Auto generate 70/30 split

Search training set sample documents

Mortgage Agreement3.pdf
Mortgage Agreement4.pdf
Mortgage Agreement5.pdf

Search test set sample documents

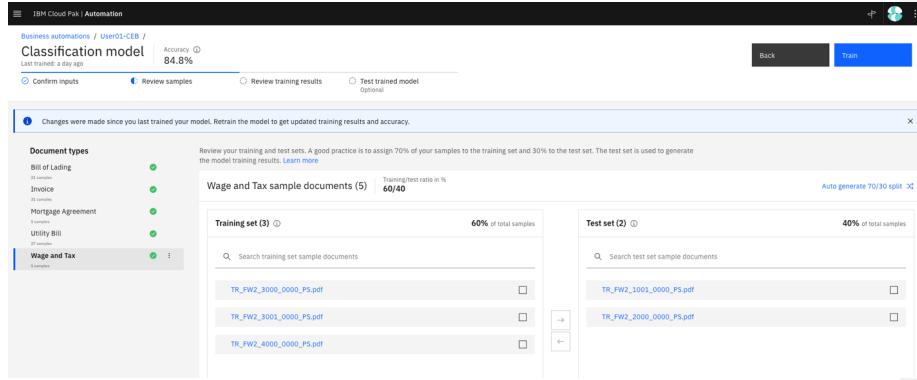
Mortgage Agreement1.pdf
Mortgage Agreement2.pdf

- _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



The screenshot shows the 'Classification model' page in the IBM Cloud Pak for Automation interface. At the top, it displays 'Accuracy: 84.8%' and a note that it was 'Last trained a day ago'. Below this are four buttons: 'Confirm inputs', 'Review samples', 'Review training results', and 'Test trained model (optional)'. A message box at the top indicates that changes were made since the last train, prompting the user to retrain for updated results. The main area is titled 'Wage and Tax sample documents (5)' and shows a split between a 'Training set (3)' (60% of total samples) containing three files: TR_FW2_3000_0000_P5.pdf, TR_FW2_3001_0000_P5.pdf, and TR_FW2_4000_0000_P5.pdf; and a 'Test set (2)' (40% of total samples) containing three files: TR_FW2_1001_0000_P5.pdf, TR_FW2_2000_0000_P5.pdf, and TR_FW2_3000_0000_P5.pdf. There are also search bars for both sets.

- _6. Click on TRAIN to launch the training. This may take a several minutes. You will see a progress bar has training progresses.



The screenshot shows the 'Classification model' page during the training process. The progress bar at the top indicates '30% complete' and 'About 21 minutes remaining'. The other buttons ('Confirm inputs', 'Review samples', 'Review training results', 'Test trained model (optional)') are visible below the progress bar.

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, it displays 'Classification model' with an accuracy of 96.9%. Below this, there are tabs for 'Confirm inputs', 'Review samples', 'Review training results' (which is selected), and 'Test trained model'. A message box says 'Model trained successfully!' and 'Accuracy has been updated to reflect the latest changes.' On the left, a sidebar lists 'Document types' including 'Bill of Lading', 'Invoice', 'Mortgage Agreement', 'Utility Bill', and 'Wage and Tax'. The main area shows 'Training results' with a table:

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

- _7. Click on each of the document types. Notice the confidence levels. The both the Mortgage Agreement and Wage and Tax have a confidence of low. 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

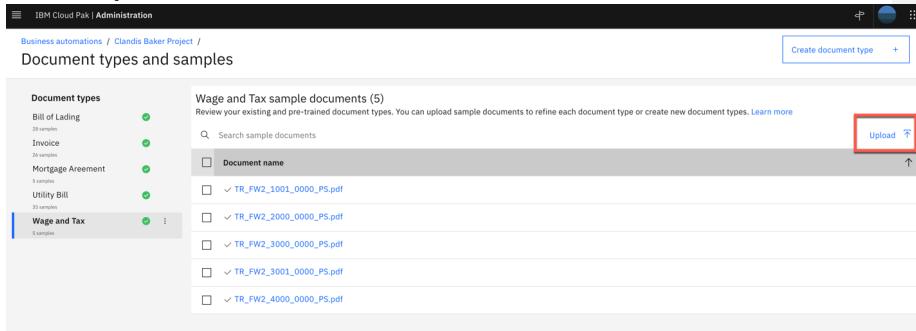
7.1 How do I improve my results?

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 *Group 2 - Design Classification docs*.

_5. Click on Build tab then let's 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.

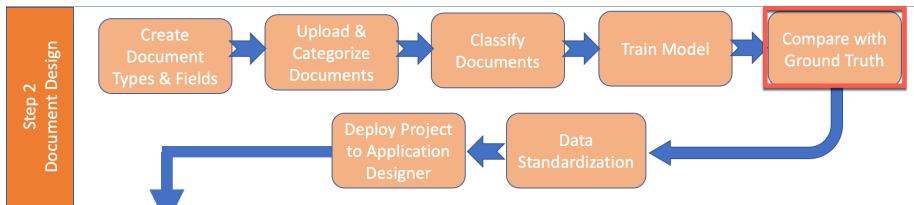
_10. Now look at the confidence score for **Wage and Tax**.

_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.

Automation Document Processing Lab

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

The screenshot shows the 'Extraction model' step in the 'Business automation / Carlos Baker Project'. It displays two lists: 'Training set (14)' and 'Test set (6)'. Both lists show PDF files under the 'Wage and Tax sample documents' category. A yellow warning box at the top states: '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. Learn more'.

_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 'Business automation / User01-CEB'. It displays a list of 'Wage and Tax sample documents (3)'. The 'Teach samples' button is highlighted with a red box. A yellow warning box at the top states: 'Please make sure you have at least 1 reviewed document to train the model. Annotate each of these documents to teach the model how to extract the fields you added in the previous step.'



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

- _ 6. We will now review the fields that were extracted, correct any that may be wrong and add others.

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

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

- Document:** TR_FW2_1001_0000_PS.pdf
- Fields Extracted:**
 - Employee's social security number: 577-22-3048
 - Employer identification number: 14-023205
 - Employee name, address, and ZIP code: Michael Robert David Schubert, 56334 Full Street Avenue Unit 1234, Minneapolis, Minnesota 55411-1234
 - Address: 123456 A7B
 - Employee first name and initial: Michael
 - Employee last name: Robert
 - Employee middle name: David
 - Employee suffix: Schubert
 - Employee address and ZIP code: MN 123456789
 - State wages, tips, etc.: 123456789.99
 - State income tax: 123456789.99
 - Local wages, tips, etc.: 123456789.99
 - Local income tax: 123456789.99
 - Identify items: ABCDEFGH
- Extracted Fields Table:**

Field Name	Value Captured
Federal Income Tax Withheld	123456789.99
Social Security Wage	123456789.99
Medicare Wage	123456789.99
Social Security Tip	123456789.99
Medicare Tip	123456789.99
Social Security	123456789.99
Medicare	123456789.99
Alimony	123456789.99
Dependent care benefits	123456789.99
Nonqualified plan	123456789.99
Other	123456789.99
AAA 888 CCCC 12345678.90	123456789.99
AAA 888 CCCC 12345678.90	123456789.99
AA	123456789.99
- Actions:**
 - Show detected fields
 - Keyboard shortcuts on
 - Sort by: Date created
 - Save selection
 - Pending aliases | View all aliases (3)
 - None (0)



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).

- 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 document titled 'TR_FW2_1000_0001F.pdf'. The document is a W-2 form for the year 2020. The 'Federal Income Tax Withheld' field is highlighted with a red box. A floating panel titled 'Value Captured' lists the extracted value '1800.00' under the field name 'Federal Income Tax Withheld'. The 'Save selection' button is also highlighted with a red box.

- ADP was able to find the field and will ask if you want to save match of value captured along with the field label. Select Save Selection

Notice a green check mark signifies this field is complete.

The screenshot shows the IBM Cloud Pak Administration interface with a document titled 'TR_FW2_1001_0000_PS.pdf'. The 'Federal Income Tax Withheld' field is marked with a green checkmark, indicating it is completed. A floating panel titled 'Value Captured' shows the extracted value '123456789.99' under the field name 'Federal Income Tax Withheld'. The 'Saved!' button is highlighted with a red box.

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

- _3. Moving to Employee Name and Address field.. Here it did pick up the address but missed the name. **Click on the Dismiss button("Employee's first name and initial"). Again, Click on Save match**

The screenshot shows the ADP software interface with the W-2 Wage and Tax Statement for 2020. On the left, the W-2 form is displayed with various fields filled out. On the right, a 'Recommended matches' overlay is open, listing suggested addresses based on the input '4326 Aldrich Rd Minneapolis, MN 55412'. The top suggestion is highlighted with a green checkmark. Below the suggestions, there are buttons for 'Edit selection', 'Dismiss', and 'Seeing duplicates?'. At the bottom of the overlay, there are buttons for 'Capture subfields', 'Pending aliases', and 'Mark this document as ready for training.'

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

- _4. You will see that there are a series of blue underlines below all the characters found. We are interested in getting the "Employee's First Name" data and the field value. **Click on the Draw button under Field value. Using your mouse select the Field Label then Click on the Draw button under Field Value and select Name and address (green box), then Select Save selection**

Automation Document Processing Lab

The screenshot shows the IBM Cloud Pak Administration interface. On the left is a PDF of a W-2 form for the year 2020. The form includes fields for employer information, employee name, and various tax-related amounts. On the right is a 'Field Name' and 'Value Captured' table. A specific row for 'Employee Name and Address' is highlighted. Two boxes are drawn around the 'Last name' field in the W-2 form and the corresponding 'Employee Name and Address' field in the table. A red box highlights the 'Save selection' button at the bottom of the table interface.

- _5. For the Employee Social Security field if it looks good, **Click on Save selection**.
- _6. Continue to process for the remaining fields, using either method as described above, clicking on the Save selection if correct or Dismiss and use blue lines if Key Value Pair (KVP) is correct or drawing a box around needed label or value.
- _7. Once complete **check the box** next to “Mark this document as ready for training” at the bottom

This screenshot is similar to the one above, showing the W-2 form and the field matching tool. However, a red arrow points to the 'Mark this document as ready for training' checkbox located at the bottom of the right-hand interface. The 'Save selection' button is also visible here.

- _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 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.

Note: When completing the remaining documents, you may run across one where there are nothing but blue dotted lines. These blue lines are showing the characters that ADP did pick up. For example:

The screenshot shows two side-by-side W-2 Wage and Tax Statement forms from 2019. The left form is the original document, and the right form is the extracted data with blue underlines indicating captured fields. A floating window on the right displays the extracted data in a grid format, allowing users to draw boxes over specific fields to update both the label and the value.

Field Name	Value Captured
Federal Income Tax Withheld	123456789.99
State Income Tax Withheld	123456789.99
Local Income Tax Withheld	123456789.99
Employer Identification Number (EIN)	123456789.99
Employee Name and Address	Michael Robert David Smithson III 56334 Full Sized Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
State Name	MN
State Tax ID Number	123456789
State Wage, Tax etc. No.	123456789.99
Local Wage, Tax etc. No.	123456789.99
Local Income Tax No.	123456789.99
Local Income Tax Type	A B C D E F G H
Other	123456789.99
Employer Name, Address, and ZIP Code	Long Lengthy Name The Corporation 56334 Full Sized Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
Employee Name, Address, and ZIP Code	Michael Robert David Smithson III 56334 Full Sized Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
Copy B To Be Filed With Employee's FEDERAL Tax Return	This information is being furnished to the Internal Revenue Service. Dept. of the Treasury - IRS Off. No. 046-008
Copy B To Be Filed With Employee's FEDERAL Tax Return	This information is being furnished to the Internal Revenue Service. Dept. of the Treasury - IRS Off. No. 046-008

By simply clicking on the Field value it will populate both the label and field value with a pop up window asking if you want to save the match.

Automation Document Processing Lab

_10. Once complete review of all the sample documents **Click on the Back link**

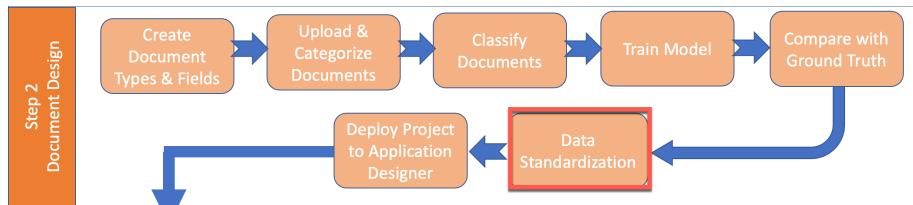
8.2 Train extraction model

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



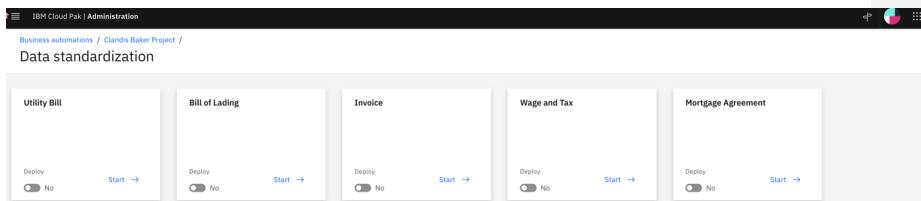
Next, we may need to standardize the data that will be presented in the user interface and how it will be stored in the FileNet repository for example. Data standardization is the process of defining attributes for a data field in a standardized way. This is done using data definitions. These definitions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

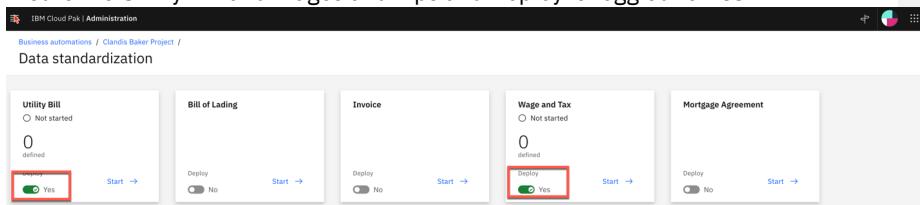
Category	Status	Count	Details
Document types and samples	Ready	4	samples on average
Classification model	Ready	3	types trained 100% accuracy
Extraction model	Ready	3	types trained 97% accuracy
Data standardization	Not ready		Map fields to new or existing data definitions.

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

Automation Document Processing Lab



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



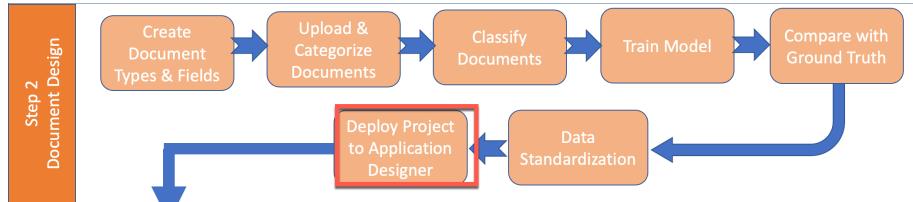
_3. Click on Start on either selected deployment.

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 /](#)

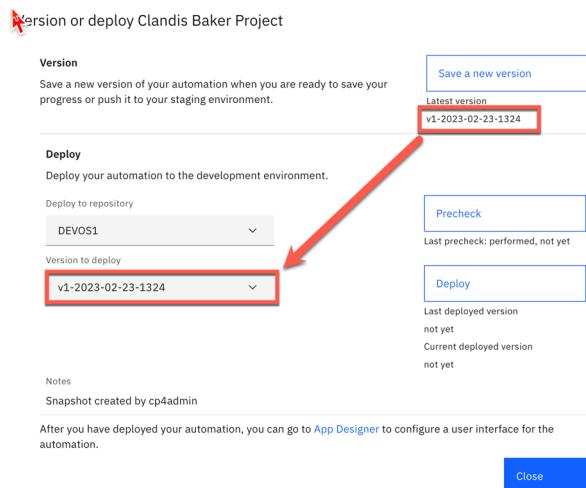
10 Version and deploy your project



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

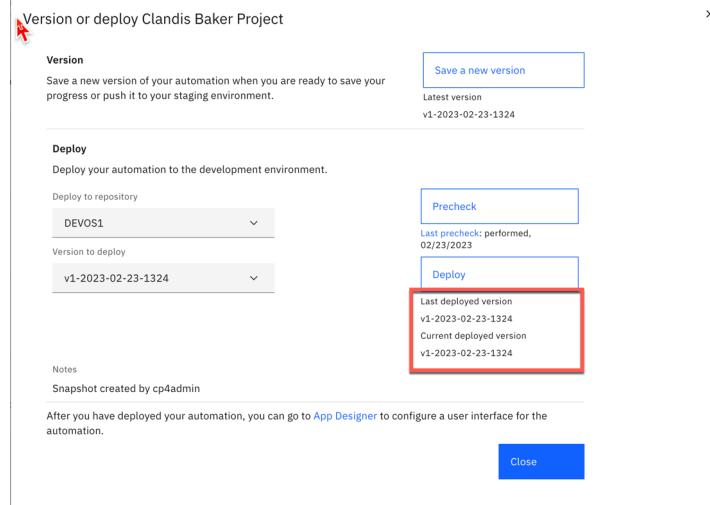


... also, in the top corner has the “Latest Version”

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.

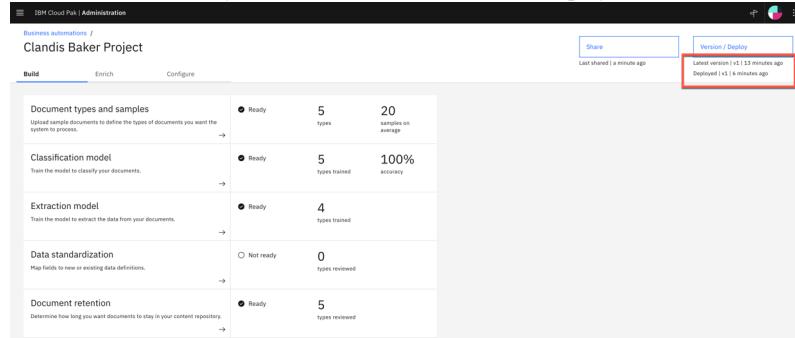


Note that you do not have to remain in the deploy screen while it is versioning or deploying. You can always click the button and then go back into any other screen if you like. It will run in the background. If you do this, just keep an eye on the top right of your screen for deployment status.

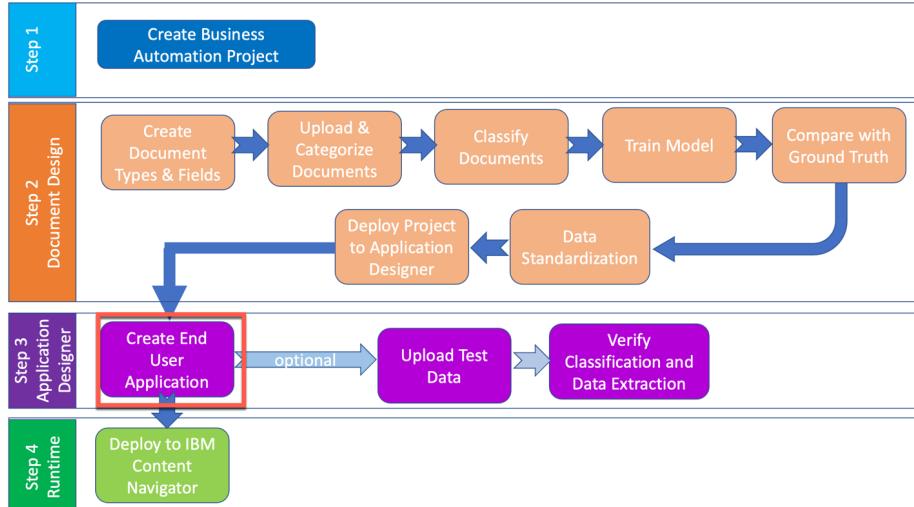
- _5. Click **Close** button.

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.

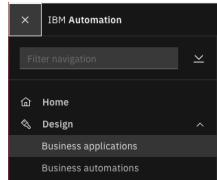
11.1 Create your Runtime Application.

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



and **selecting Business Applications**

Automation Document Processing Lab



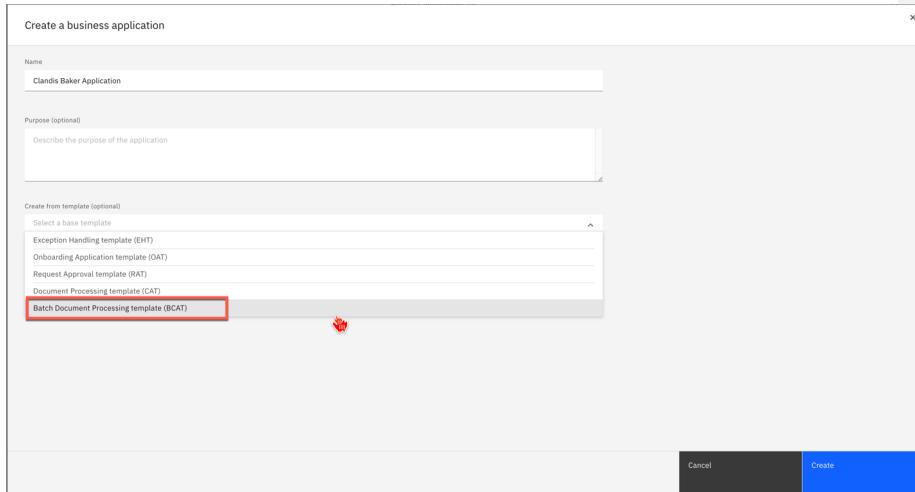
_2. From the **Create** drop down list, select Application

A screenshot of the IBM Cloud Pak Administration interface under 'Business applications'. The left sidebar shows 'Create' (selected), 'Import', 'Application' (highlighted in red), 'Template', and 'Toolkits'. The main area shows a message: 'There are no applications to display...yet.' with a link to 'Create'. Below this are three template cards: 'Request Approval template', 'Onboarding Application template', and 'Exception Handling template', each with a 'Last updated' date of 02/20/2023.

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

A screenshot of the 'Create a business application' dialog box. The 'Name' field contains 'Clandis Baker Application' (highlighted with a red box). The 'Purpose (optional)' field is empty. Under 'Create from template (optional)', the 'Select a base template' dropdown is open, showing options like 'Exception Handling template (EHT)', 'Onboarding Application template (OAT)', etc. At the bottom are 'Cancel' and 'Create' buttons, with 'Create' being highlighted in blue.

- _4. In the Create Form Template 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**

Create an application

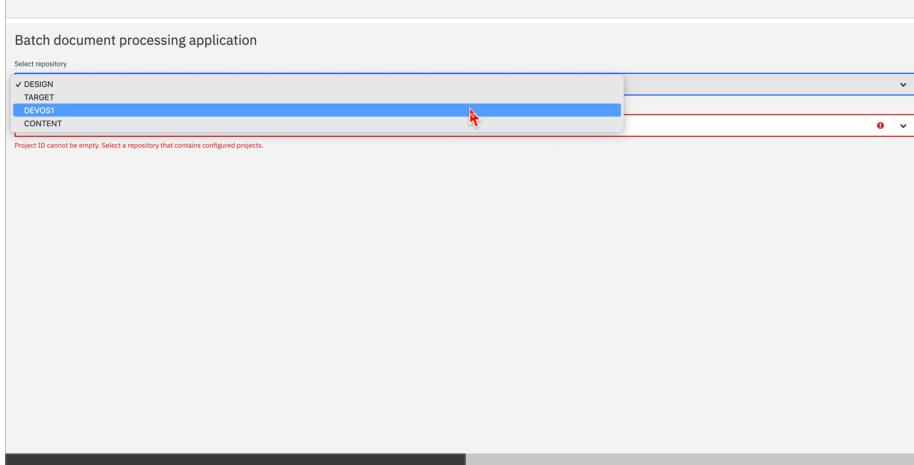
Batch document processing application

Select repository

✓ DESIGN
TARGET
DEVOS1
CONTENT

Project ID cannot be empty. Select a repository that contains configured projects.

Back Create



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

Create an application

Batch document processing application

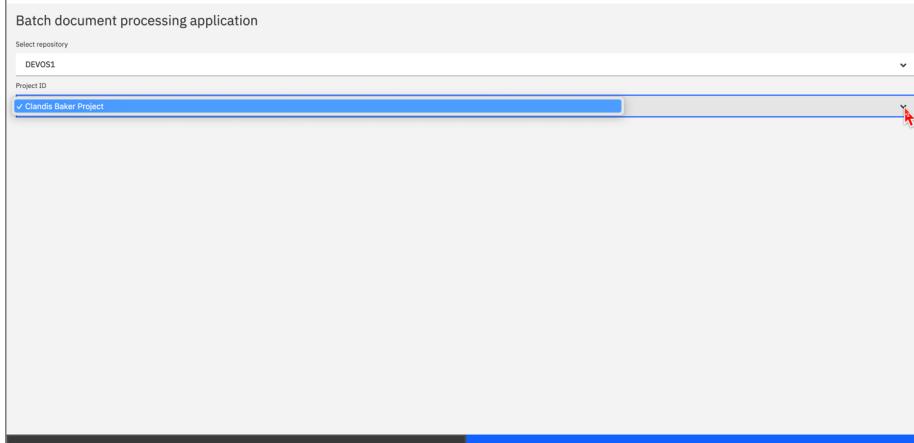
Select repository

DEVS01

Project ID

✓ Clandis Baker Project

Back Create

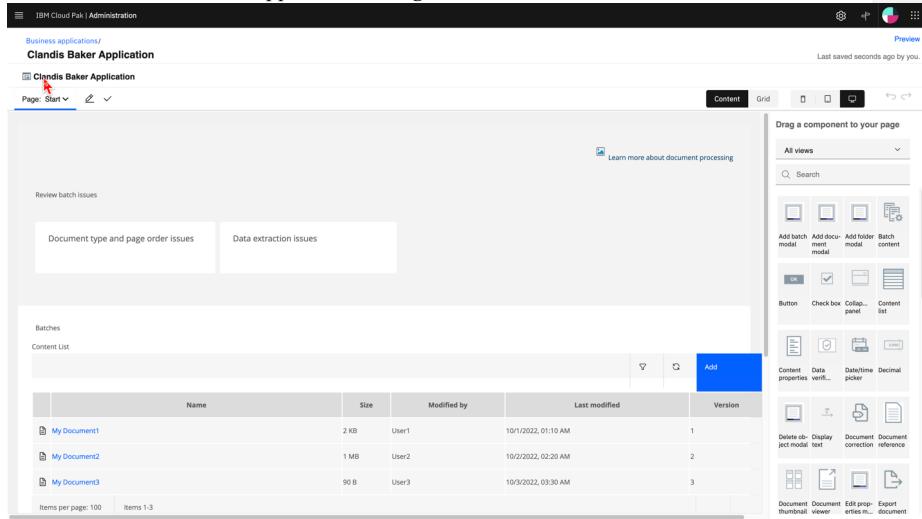


Note it may take a minute or two before this updates and you can see your project

Automation Document Processing Lab

_8. Click **Create**

You should now be in the *Application Designer*



The screenshot shows the 'Content' tab of the Application Designer. On the left, there's a sidebar with 'Business applications / Cländis Baker Application'. Below it, a 'Page: Start' dropdown and a 'Content' tab are visible. The main area has sections for 'Review batch issues', 'Document type and page order issues', and 'Data extraction issues'. To the right, a 'Drag a component to your page' panel lists various components like 'All views', 'Search', 'Add batch modal', 'Add document modal', etc. At the bottom, there's a table titled 'Batches Content List' with columns for Name, Size, Modified by, Last modified, and Version. Three items are listed: 'My Document1', 'My Document2', and 'My Document3'. An 'Add' button is highlighted with a red arrow.



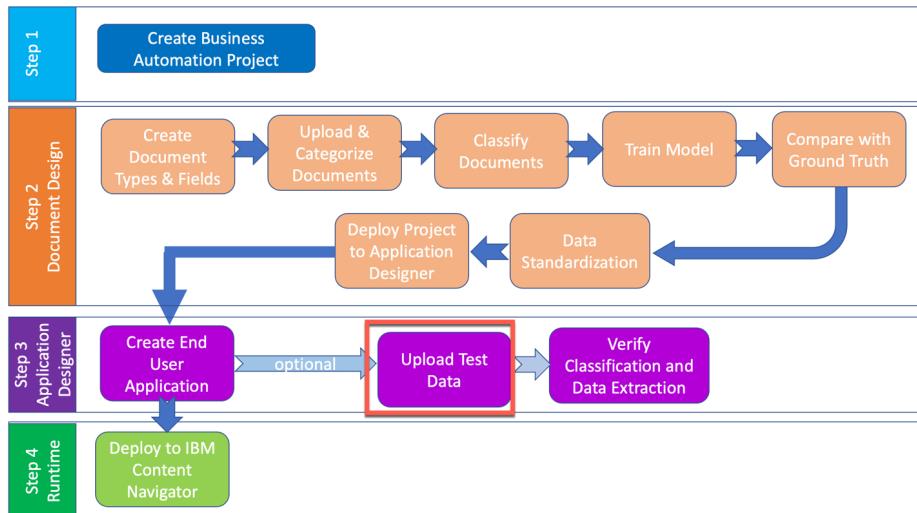
Batch Document Processing template (BCAT) has all the necessary pages and configuration to start using the application. Using this designer user interface, you have the option to further customize the application, such as its page design or actions, to fit your requirements.

_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

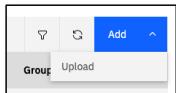


_1. You should be in the default application user interface for ADP.

The screenshot shows the 'Review batch issues' screen of the IBM Cloud Pak for Integration application. It displays two sections: 'Document type and page order issues' and 'Data extraction issues'. Both sections show 0 batches. Below this is a 'Batches' table with 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.

Upload new batch

* Display Name
Batch 1

Description

Priority
High

- _4. Click **Select files**.

Navigate to the samples folder previously downloaded and use the *Group 3 Application Runtime Set* folder documents. Select all the files in the folder

- _5. Click **Open**

You will see a window that will give the operator a chance to manually classify the documents before they are ingested. By clicking on one of the files you will be presented with an option to manually classify the documents. In the example below would be how to manually classify a document.

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.

File Name	Document Type
B_PO_5.pdf	Auto Classify
DE_FW2_1000_0001F.pdf	Auto Classify
DE_FW2_4000_0011F.pdf	Auto Classify
DE_FW2_4001_0001S.pdf	Auto Classify
DE_FW2_4001_0010F.pdf	Auto Classify

1 items selected

Classify | Auto Classify | Deselect

Cancel Add

We are not going to do this but instead let ADP auto classify them.

Automation Document Processing Lab

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.

The screenshot shows a table titled 'Filter List' with columns for 'File Name' and 'Document Type'. There are five rows, each with a checkbox in the first column. The 'Document Type' column contains the value 'Auto Classify' for all rows. The last row is 'DE_FW2_4001_0001S.pdf'. At the bottom of the table are 'Cancel' and 'Add' buttons.

	File Name	Document Type
<input type="checkbox"/>	B_PO_5.pdf	Auto Classify
<input type="checkbox"/>	DE_FW2_1000_0001F.pdf	Auto Classify
<input type="checkbox"/>	DE_FW2_4000_0011F.pdf	Auto Classify
<input type="checkbox"/>	DE_FW2_4001_0001S.pdf	Auto Classify
<input type="checkbox"/>	DE_FW2_4001_0010F.pdf	Auto Classify

Cancel Add

_6. Click on the Add button.

The screenshot shows the 'Review batch issues' section with two boxes: 'Document type and page order issues' (0 batches) and 'Data extraction issues' (0 batches). Below is the 'Batches' section with a table. A progress bar at the top of the table indicates '3 of 5 files processed'. A red arrow points to the 'Add' button at the top right of the table header.

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

Review batch issues

Document type and page order issues 0 batches

Data extraction issues 0 batches

Learn more about document processing

Batches

Items per page: 100 1-1 of 1 items

Add

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 shows the same 'Review batch issues' and 'Batches' sections as the previous step. A red arrow points to the three-dot menu icon at the end of the table header row.

Name	Files	Priority	Status	Added on	Added by	Group	Location
Batch01	5	High	<div style="width: 100%;">Documents uploaded</div>	02/23/2023, 10:49 AM	cp4admin		

Review batch issues

Document type and page order issues 0 batches

Data extraction issues 0 batches

Learn more about document processing

Batches

Items per page: 100 1-1 of 1 items

_8. Click Submit

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

The screenshot shows a dashboard titled "Review batch issues". It contains two tiles: "Document type and page order issues" (1 batches) and "Data extraction issues" (0 batches). Below these is a table titled "Batches" with one row:

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

At the bottom left, there are buttons for "Items per page: 100" and "1-1 of 1 items".

11.3 Correct any classification errors.

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

The screenshot shows a list view titled "Document type and page order issues". It has a header row with columns: Name, Priority, Status, Added on, Added by, Group, and Location. There is one item listed:

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

At the bottom left, there are buttons for "Items per page: 100" and "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.

Automation Document Processing Lab

Batch1

The screenshot shows a list of documents in a 'Batch1' folder. One document, 'B_PO_5.pdf', is highlighted with a red border. To its right is a detailed view of a 'PURCHASE ORDER' from 'RUBE'S Meat Co.' to 'Chicken Run Ranch'. The order details include 200 pieces of item #61 (Whole Chicken) at £1.00 each, totaling £340.00, and 150 packs of item #62 (One Day Old Chick) at £1.00 each, totaling £157.50. The total amount is £602.50.

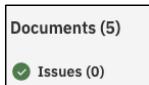
Commented [CB1]: May need new screen

- _3. Most of the document types are correct but it looks like a PO got mixed into our batch so we can **Click** on the **Trash can** to delete it from the batch. And **select OK** to delete it.

Batch1

The screenshot shows the same batch of documents. The purchase order document 'B_PO_5.pdf' now has a red-bordered trash can icon next to it, indicating it is selected for deletion. The rest of the interface remains the same, displaying the list of documents and the detailed view of the purchase order.

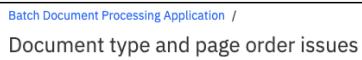
- _4. Review all documents to ensure everything is correct. If the system no longer detects any issues, you should see a green checkmark near the top of the document list.



_5. **Click Save Changes and Submit** to save your changes and have the batch processed.

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

_6. **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.**



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

Automation Document Processing Lab



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

A screenshot of a web application interface titled 'Batch Document Processing Application / Batches with data extraction issues / Batch1'. The page shows a table of documents with their status and modification details. One document, 'DE_FW2_4001_0010F.pdf', is highlighted with a red border and has a yellow triangle icon next to it, indicating it has data issues. The table includes columns for Name, Issues, Status, Modified on, and Modified by. A 'Submit' button is visible at the top right. At the bottom, there is a message 'Items per page: 100 1-4 of 4 items'.

Notice one of the documents has Data issues.

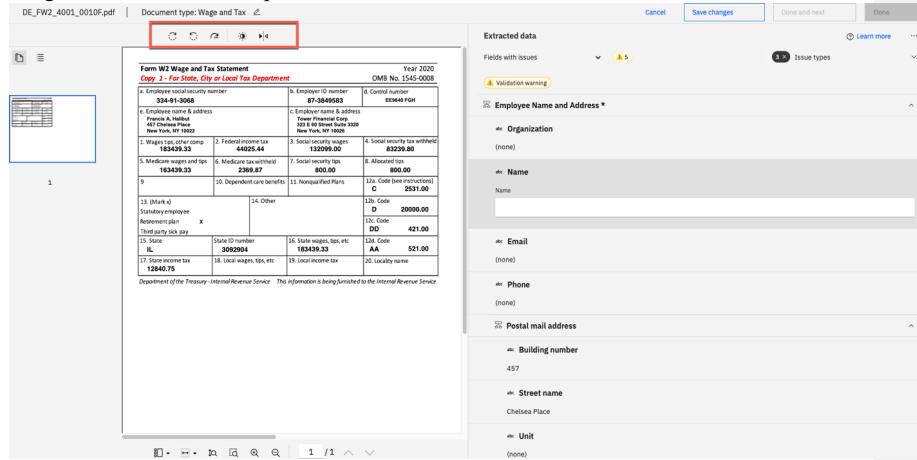
- _3. Click on the document to open it. Zoom in a bit to get a better picture of the document.

A screenshot of a tax form viewer for 'Form W2 Wage and Tax Statement'. The form is displayed on the left, and a detailed view of the extracted data is shown on the right. A red arrow points to the magnifying glass icon in the bottom right corner of the viewer window, which is part of a toolbar. The right panel shows various fields like Employee Name and Address, Organization, Name, Email, Phone, and Postal mail address, each with a '(none)' placeholder. A validation warning is visible at the top of the right panel.

Take a moment to discover the image viewer features.

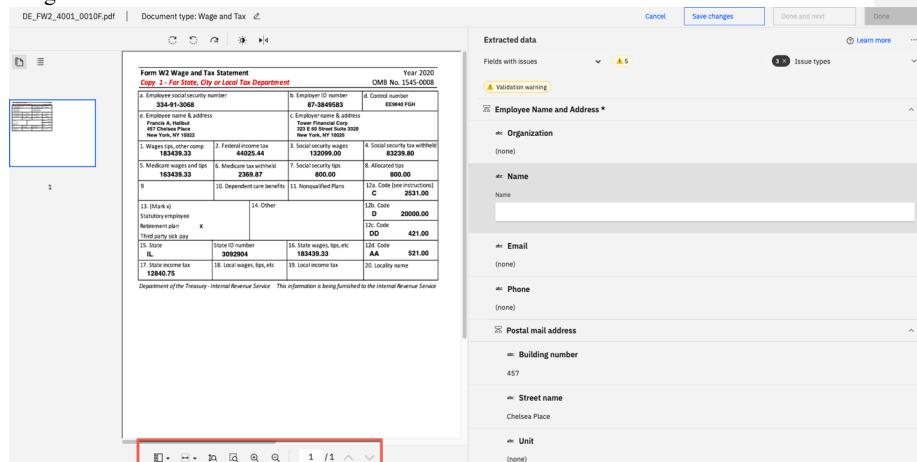
Automation Document Processing Lab

Image viewer features at top:



- Rotate image.
 - Visual effect adjustment
 - Invert

Image viewer features at bottom:



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

Automation Document Processing Lab

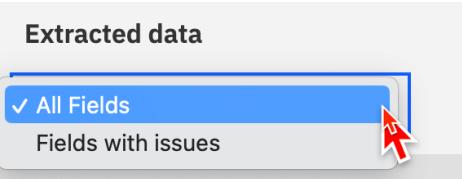
Field features

The screenshot shows a document titled "Form W-2 Wage and Tax Statement" from the Internal Revenue Service. The document includes fields for Employee Social Security number (338-91-3066), Employer ID number (87-3849583), Control number (EE2640 FGH), and various wage and tax amounts. To the right, a "Extracted data" panel is open, showing a dropdown menu with "Fields with issues" selected. The "Employee Name and Address" section is expanded, showing fields for Organization (none), Name (Name), Email (Email), Phone (Phone), and Postal mail address (Building number 457, Street name Chelsea Place, Unit none). A validation warning icon is visible next to the "Employee Name and Address" section.

- Show all fields.
- Show fields with issues.

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

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



_5. Click on the ALL Fields.

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

Change the Extracted data back to Fields with issues.



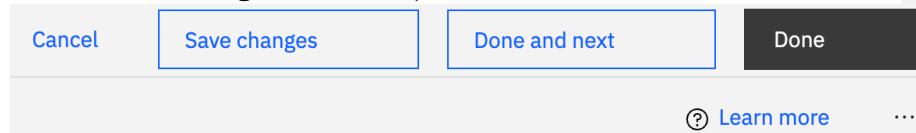
The Employee Name and Address is a mandatory field. It is also a composite field. Remember when we defined this field we picked Address Information, because we wanted not only the Address block but also the name.

This is why you see multiple fields under the Employee Name

_6. Click on Name and with your mouse select the person's name.

The screenshot shows a document processing application interface. On the left is a preview of a W2 form for Year 2020, page 1, from the City of Los Angeles Tax Department. The form includes fields for employee information, wages, taxes, and benefits. On the right is an 'Extracted data' panel. Under the 'Employee Name and Address' section, there is a 'Name' field containing 'Francis Hallbut'. A red arrow points from the text 'Click on Name and with your mouse select the person's name.' to this 'Name' field in the extracted data panel.

_7. Click on Save Changes box at the top.



_8. For the remaining fields there are no extraction issue only low confidence characters. Click on Dismiss for each field with a yellow validation warning.

_9. Click on Done and next..

_10. All documents have been processed Click on Submit at the top to complete the batch.

END OF LABS

12 Export Import Project.

From the Business Automations

- _1. From the Business Automations screen **select Document Processing**.

The screenshot shows the 'Business automations' section of the IBM Cloud Pak Administration interface. On the left, there's a sidebar with 'Create' and 'Import' buttons. Below them is a list of automation types: 'Published automation services' (with a right arrow), 'Decision' (with a right arrow), 'Document processing' (which is highlighted with a blue border and has a right arrow), 'Workflow' (with a right arrow), and 'External' (with a right arrow). The main area displays a single project named 'Clandis Baker Project' with the note 'Last edited 02/23/2023'.

- _2. Select <your project name> Click open

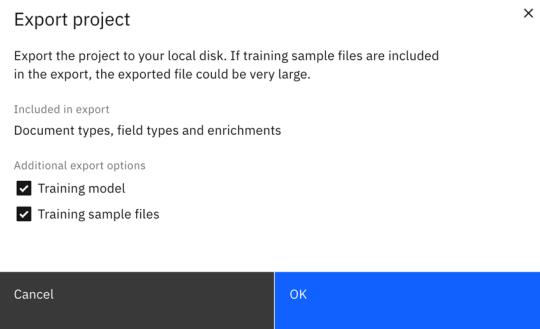
This screenshot is similar to the previous one, showing the 'Business automations' screen. The 'Document processing' category under 'Decision' is selected. A red box highlights the 'Open' button next to the project name 'Clandis Baker Project'.

- _3. From the Main screen select the Configure tab

This screenshot shows the 'Configure' tab for the 'Clandis Baker Project'. The 'Configure' tab is highlighted with a blue underline. On the left, there are sections for 'Import / Export ontology' (with 'Language settings' and 'Git server configuration' buttons), 'Export project' (with a 'Export project' button), and 'Import project' (with a 'Import project' button). At the top right, there are 'Share' and 'Version / Deploy' buttons. Below the tabs, there are notes about sharing and deployment: 'Last shared 12 hours ago' and 'Latest version v2 12 hours ago 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.

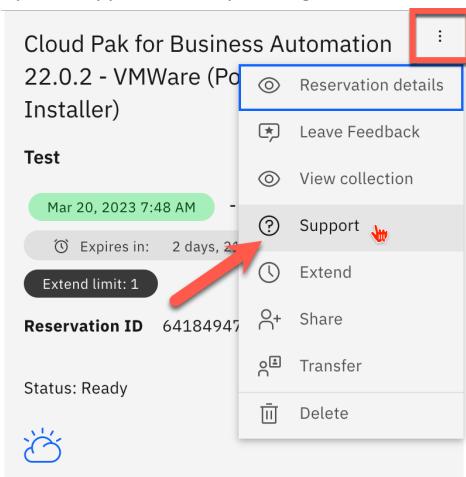
An operator fails to install and continuously shows Pending status.

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) MainMenu
#? 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:

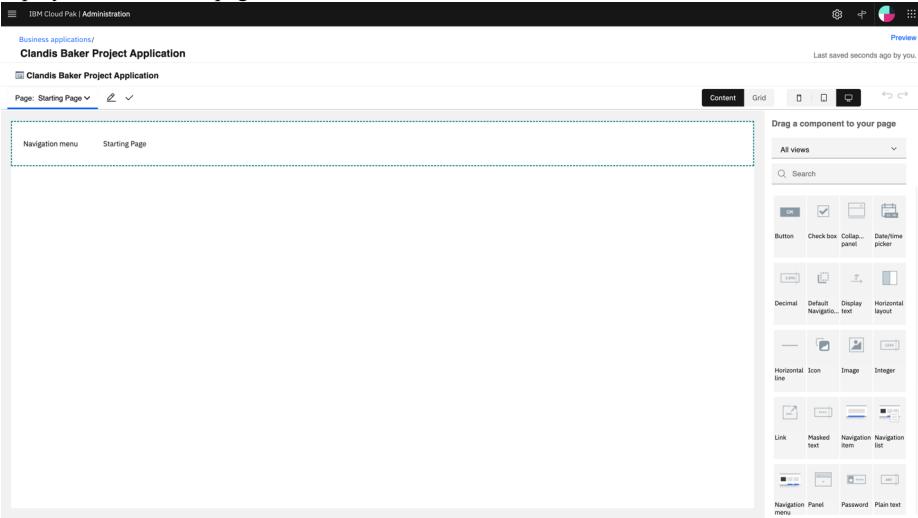
- Open the **OpenShift web console** in a browser.
- Click on **Workloads -> ConfigMaps** on the left-hand side navigator.
- Type ***access-info*** in the field next to 'Name'.

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

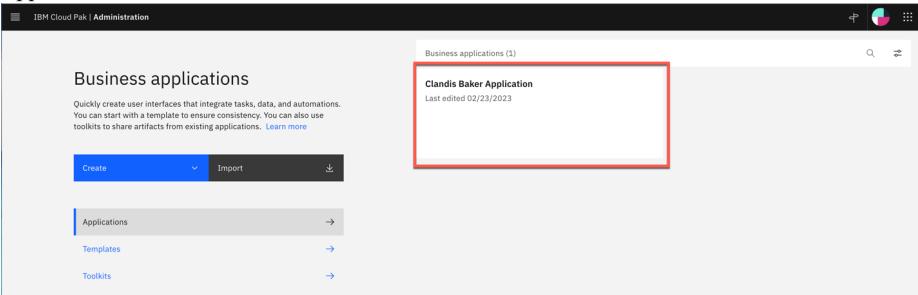
Automation Document Processing Lab

APPLICATION BLANK

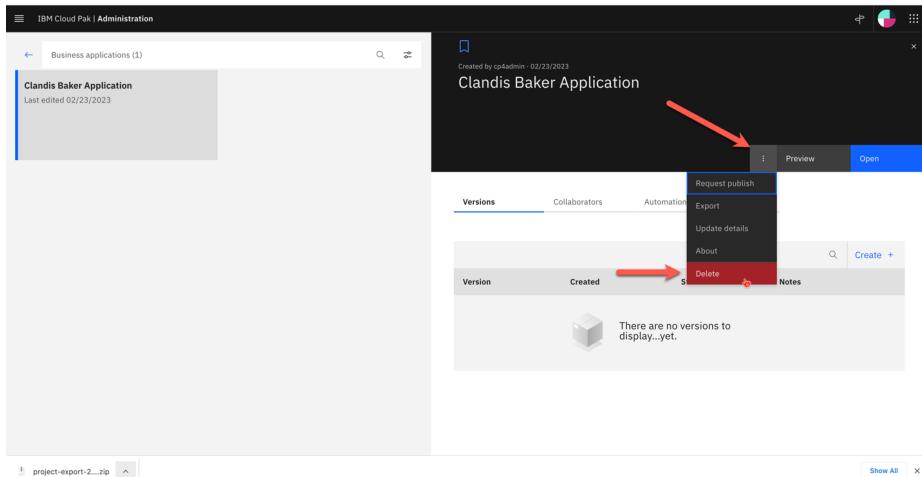
During creating of Business Application setup, sometimes on first time after project has been deployed. The Starter page is blank.



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.

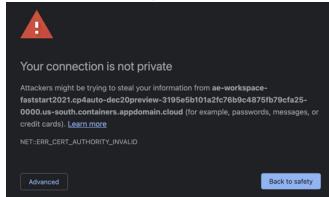
OPENING AN INCOGNITO WINDOW

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:

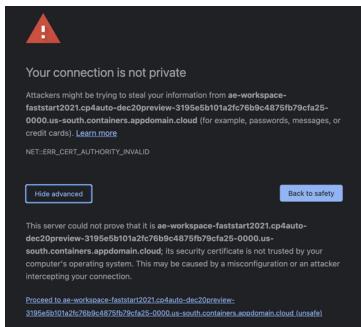
Open the Generate Security Tokens Box note and click all 3 of the links listed. This will reset the self-signed security certificates.

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

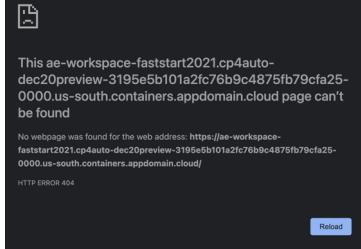


Click Advanced, and the browser window will look something like this:

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>