

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	5
1.1 Getting HELP during the lab.....	5
1.2 Abstract	5
1.3 Introduction	5
2 Getting started	7
2.1 IBM TechZone – Reserve the environment.....	7
2.1.1 Credentials	8
2.2 Set up WireGuard VPN	10
2.3 Open your IBM Cloud Environment.....	11
3 Lab Overview.....	14
3.1 How does ADP work?.....	14
4 Create Document Processing Project.....	16
4.1 Reviewing the interface.	20
4.1.1 Build Tab	20
4.1.2 Enrich Tab.....	21
4.1.3 Configure Tab	22
4.2 Configure a Wage and Tax document type.....	24
4.3 Create Wage and Tax document type.	24
4.4 Create Field	26
4.5 Create the Employee Name Address field.	28
4.6 Create Employee Social Security Number Field	29
5 Document Types and Samples Overview	33
5.1 Categorize documents.	34
6 Train classification	41
6.1 How do I improve my results?.....	45
7 Data extraction.....	47
7.1 Correcting extracted values.....	49
7.2 Train extraction model.....	54
8 Data standardization	55
9 Version and deploy your project	57
10 Application designer	59
10.1 Create your Runtime Application.	59
10.2 Upload documents for processing	64
10.3 Correct any classification errors.....	66
10.4 Correct extraction issues.....	68
11 Export Import Project.	76
Appendix A - Troubleshooting	78

1. Overview

1.1 Getting HELP during the lab

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

1.2 Abstract

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

1.3 Introduction

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

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

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

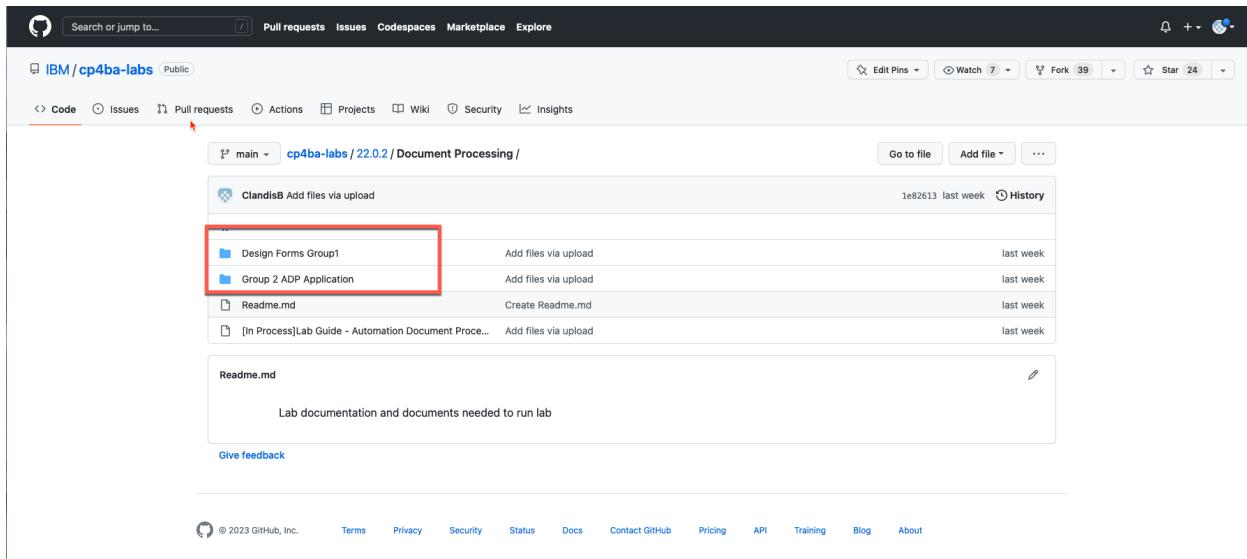
You can extract text, check boxes, forms, tables, barcodes, signature detection and even free text. With no or low code options, you can create an application that processes documents, extracts data, flags issues, and stores your documents and data. And the data enrichment capabilities ensure that the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

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

1 Getting started

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

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



The screenshot shows a GitHub repository page for 'cp4ba-labs'. The main navigation bar includes 'Pull requests', 'Issues', 'Codespaces', 'Marketplace', and 'Explore'. Below the navigation, there are tabs for 'Code', 'Issues', 'Pull requests', 'Actions', 'Projects', 'Wiki', 'Security', and 'Insights'. The 'Pull requests' tab is selected. A search bar at the top says 'Search or jump to...'. The repository path is 'main / cp4ba-labs / 22.0.2 / Document Processing /'. A file list is shown, with two files highlighted by a red box: 'Design Forms Group1' and 'Group 2 ADP Application'. Other files listed include 'Readme.md' and '[In Process]Lab Guide - Automation Document Proce...'. The bottom of the page includes links for 'Give feedback' and various GitHub footer links like 'Terms', 'Privacy', 'Security', etc.

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

1.4 IBM TechZone – Reserve the environment.

What is IBM TechZone?

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

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

1.1.1 Credentials

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

The screenshot shows the 'Pak Installer' collection page on the IBM Technology Zone. The main content area features a title 'Pak Installer: Automated environments for OpenShift, IBM Cloud Paks, and Cartridges', a 5-star rating with 40 reviews, and a detailed description of the asset. Below the description, there are tabs for 'Business value', 'Cloud Pak for Business Automation', 'Cloud Pak for Data', and 'Cloud Pak for Integration'. The left sidebar has sections for 'Authors', 'Resources', and 'Comments', with 'Authors' currently selected.

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

The screenshot shows the 'Create a reservation' page for the 'Pak Installer'. The page has a progress bar at the top with four steps: 'Select a environment/infrastructure', 'Select a reservation type', 'Fill out your reservation', and 'Complete'. The second step, 'Select a reservation type', is currently active. Below the progress bar, there are instructions for selecting a reservation type and options for 'Single environment reservation options': 'Reserve now' (selected) or 'Schedule for later'. At the bottom, there are 'Cancel', 'Reset', and 'Submit' buttons, along with a small icon. To the right of the form, there is a decorative illustration of a city skyline with a lightbulb and clouds.

- _5. Create a Reservation
Based on the reservation type you are making, provide the required information
Customer Demo : Need a short customer-facing demonstration
Practice/Self-Education: Need to gain experience
Standard proof of concept; Need an environment for a standard product use case.
Custom Proof of concept: Need a complex, customized environment.
Testing: Need to test a specific function, configuration, or customization.
_6. For this lab **Testing** will give you 3 days plus the option to extend it for another week.
Otherwise, you will need a legitimate opportunity to leverage another reservation type.

_7. For Preferred Geography (required) select your preferred data center location

Preferred Geography (required)

Choose a preferred geography

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

_8. For VPN Access choose **Enable**

VPN Access (required)

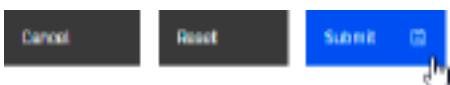
✓ Disable
Enable

_9. For Starter Service choose **docprocessing**

Starter Service (required)

✓ all
content
content-decisions
decisions
docprocessing
workflow

_10. Click Submit

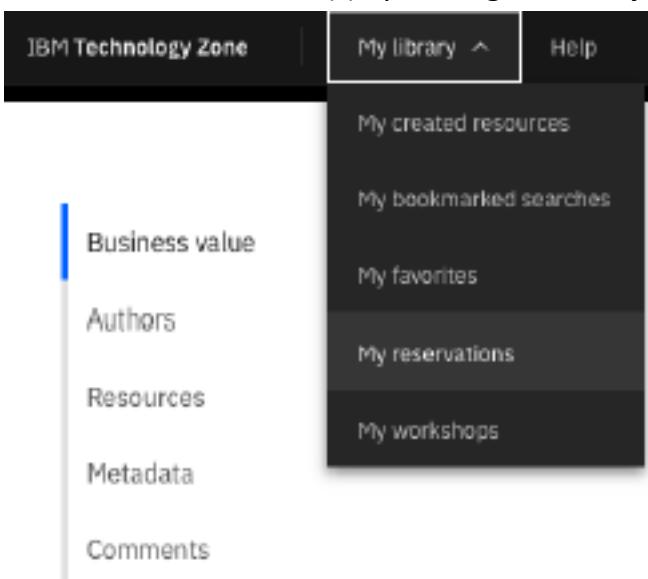


Upon receiving the Your environment is ready email, please allow up to 1 hour for the start-up services to fully complete. If after receiving email and a few hours have passed and your environment is not up, check Appendix A – Trouble Shooting for possible fix. Once the start-up process is complete you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

_11. Click My reservations



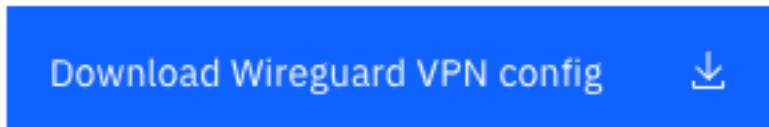
_12. Once you get the email from the IBM Technology Zone site, you can access your environment reservation(s) by clicking on the **My library** then **My Reservations**



You can also access directly using the link below
<https://techzone.ibm.com/my/reservations>

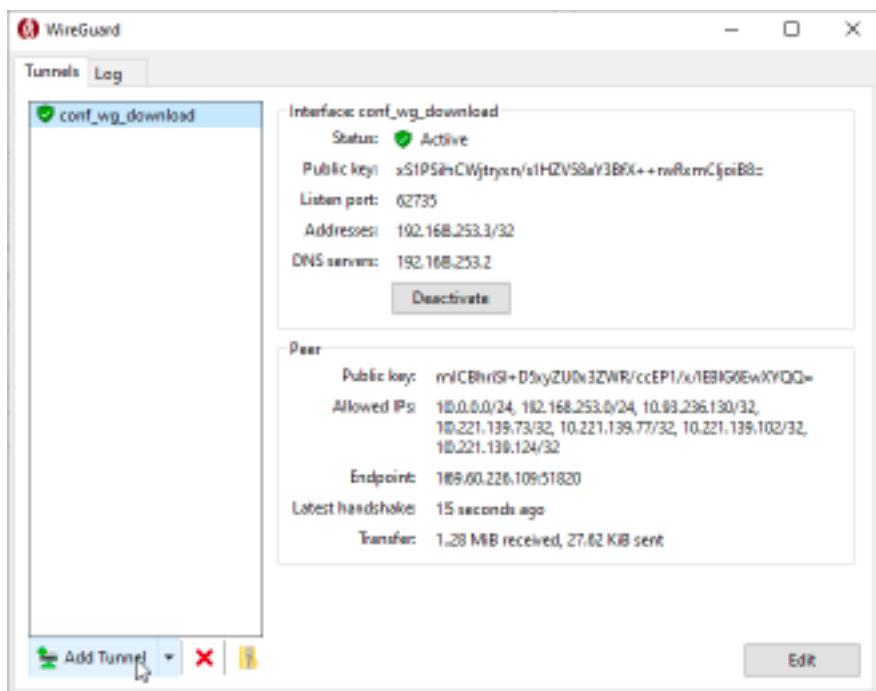
1.5 Set up WireGuard VPN

- _1. **Open** your reservation tile and scroll to bottom
- _2. **Click Download WireGuard VPN config** button to download conf_wg_download.conf to your local workstation

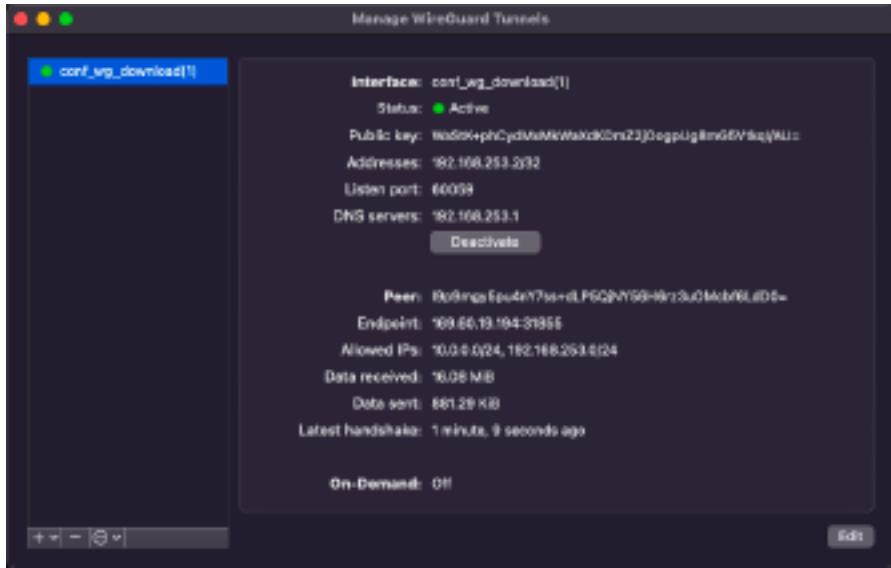


- _3. On your local workstation, install WireGuard by accessing <https://www.wireguard.com/install/>
- _4. **Launch** WireGuard
- _5. **Click Add Tunnel** and load the **conf_wg_download.conf** file.

WireGuard on Microsoft Windows



WireGuard on Mac



1.6 Open your IBM Cloud Environment

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

The screenshot shows the 'My reservations / Collection' page. A reservation for 'Cloud Pak for Business Automation 22.0.1/22.0.2 - VMWare (Powered by Pak Installer)' is listed. The status is 'Ready'. The reservation details include:

- Date: Feb 20, 2023 7:49 AM - Feb 27, 2023 7:49 AM
- Expires in: 2 days, 22 hours, 41 minutes
- Desktop
- Shared Reservation
- Purpose

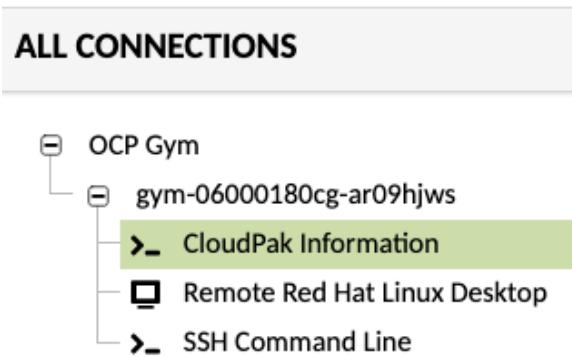
A blue button labeled 'Open your IBM Cloud environment' is highlighted. Below it, the URL is provided: <https://remote.cloud.techzone.ibm.com/guacamole/#/?username=gymuser-ar09hjws&password=zHCB8oy>.

- _2. Expand OCP Gym under All Connections

The screenshot shows the 'ALL CONNECTIONS' interface. Under the 'OCP Gym' section, the following connections are listed:

- gym-06000180cg-ar09hjws
 - CloudPak Information
 - Remote Red Hat Linux Desktop
 - SSH Command Line

_3. Select CloudPak Information



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

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

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

_5. **Enter 1** for Console information

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

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

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

Note: Controls for copy and paste in guacamole.

For Mac users:

CONTROL_OPTION_SHIFT

For Windows users:

CTRL_ALT_SHIFT

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

Cloud Pak Dashboard URL

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

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

2 Lab Overview

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

1.7 How does ADP work?

Document Processing Designer

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

Deployment tools

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

Application templates and toolkits

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

Document processing application and document management

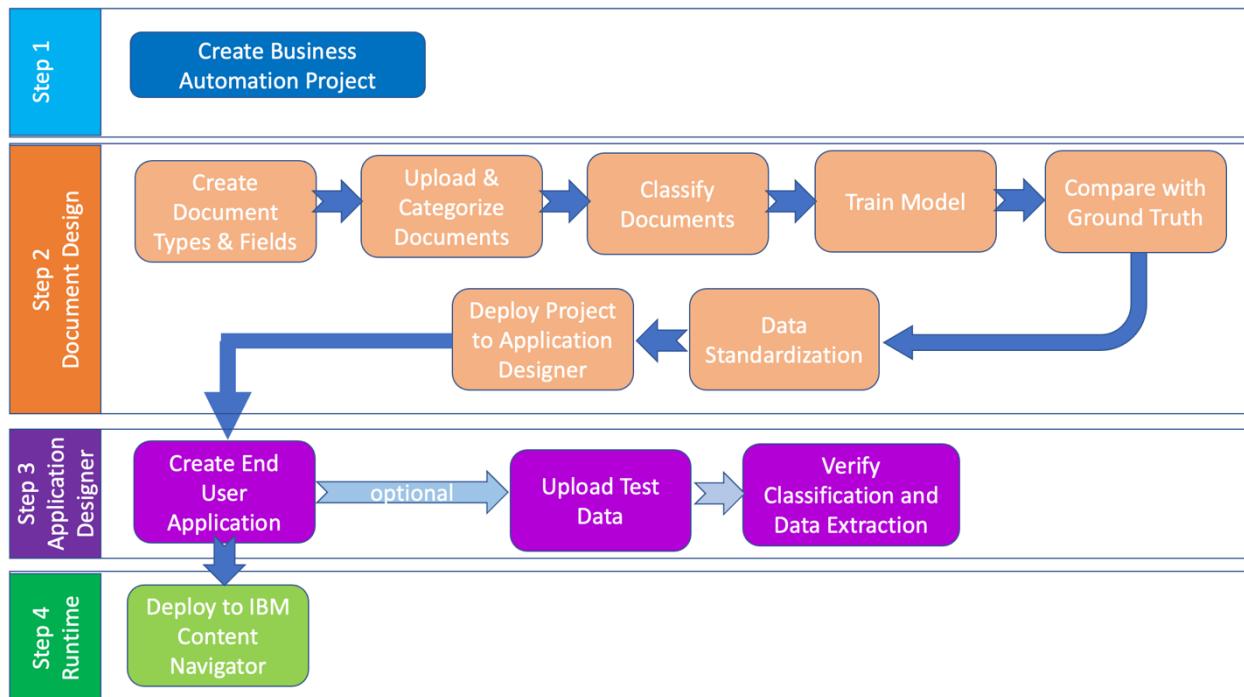
The application that you build uses the AI and deep learning to automatically detect, extract, and standardize the data in all your documents. Any anomalies are flagged according to your customized model and the priority that you set so that your document processing user can correct issues before the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

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

End result

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

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



Step 1 – Create Business Automation Project

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

Step 2 – Document Design

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

Step 3 – Application Designer

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

Step 4 – Runtime

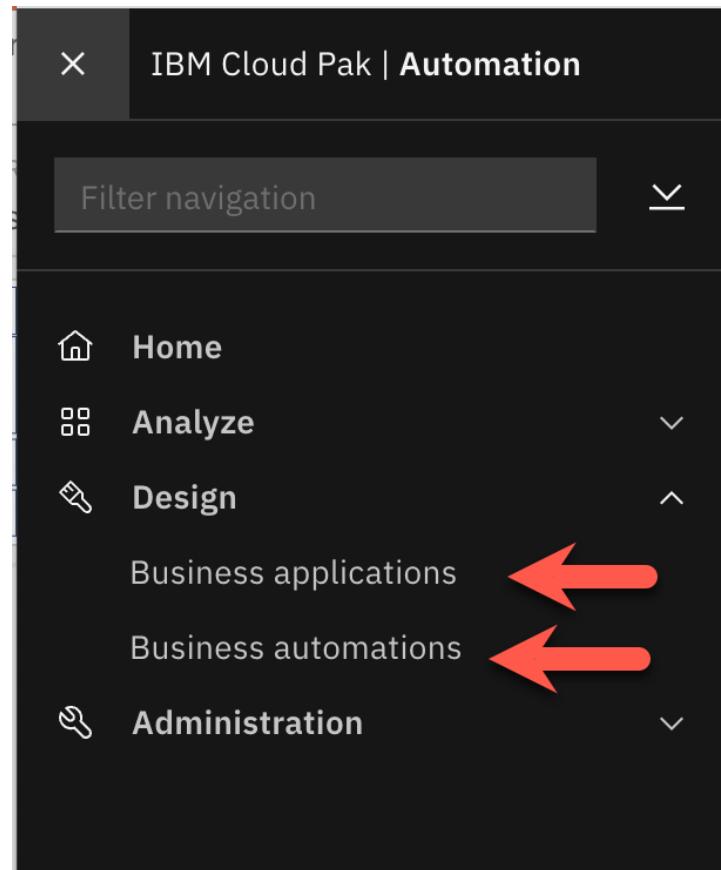
End-users would be using the runtime IBM Content Navigator interface to process documents or batches, classify document and verify extracted field data in the verification screen.

3 Create Document Processing 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.



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

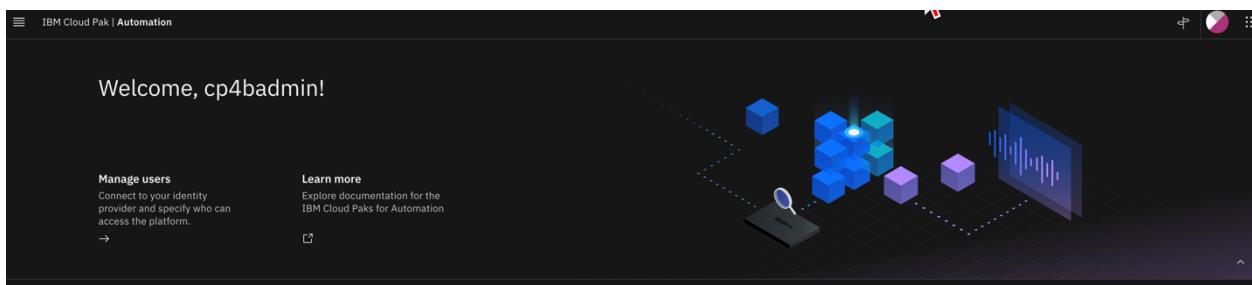
Within the Business Automations you can create or reuse automations. An automation is a collection of artifacts that fulfills a business purpose. You can publish some automation artifacts as automation services that you can call and reuse in a consistent way. Also in Business Automation, you use the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

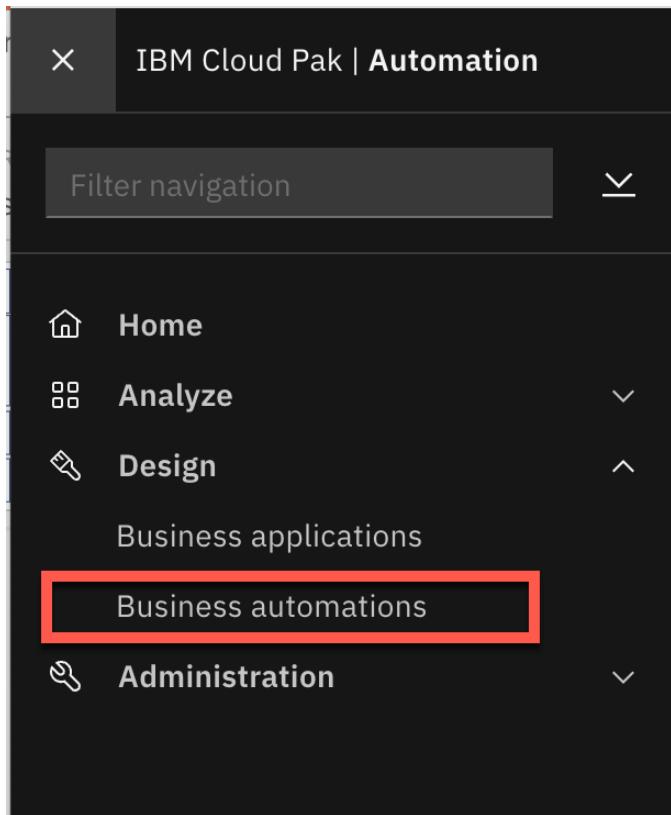
Within *Business Applications* you can quickly create user interfaces that integrate tasks, data, and automations. You can start with a template to ensure consistency. You can also use toolkits to share artifacts from existing applications.

We will start with the Business Automations.

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



- _1. Click on the hamburger menu at the top left next to IBM Automation and Click on Drop down arrow next to Design then Select Business Automations.



The following screen appears

The screenshot shows the "Business automations" screen. The title "Business automations" is at the top. A descriptive text block says: "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." Below this is a "Learn more" link. At the top right are "Create" (blue button), "Import", and a downward arrow. Below these are four categories: "Published automation services" (with an arrow icon), "Decision" (with an arrow icon), "Document processing" (with an arrow icon), "Workflow" (with an arrow icon), and "External" (with an arrow icon).

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

Business automations

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.

[Learn more](#)

The screenshot shows a user interface for managing business automations. At the top, there are two buttons: 'Create' (blue) and 'Import' (dark grey). Below these are three main categories: 'Decision automations', 'Document processing automations' (which is highlighted with a red box and has a red arrow pointing to it), and 'Workflow'. Under each category, there are sub-options: 'External' and 'Document processing' (under Document processing automations), and 'Workflow' (under Workflow). Each sub-option has a right-pointing arrow indicating further action.

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

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

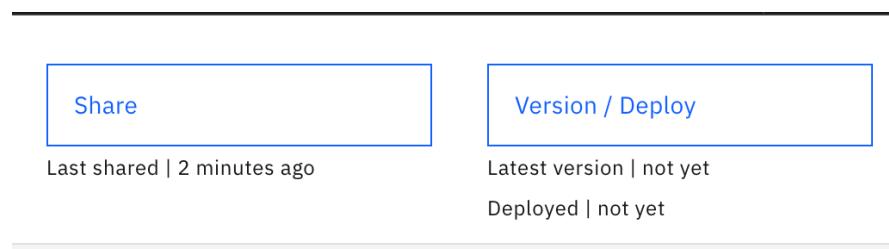
_11. Click on **Create** in the lower right-hand corner.

1.8 Reviewing the interface.

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

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.

3.1.1 Build Tab

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

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

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

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

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

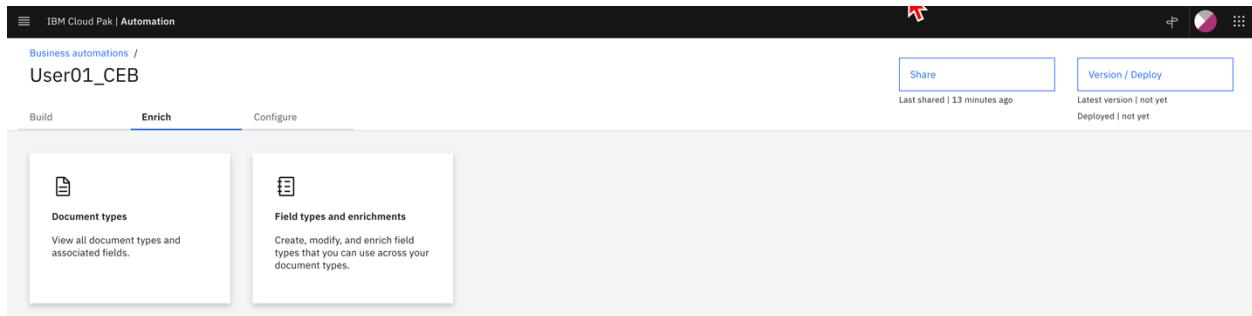
Document retention: This allows us to define how long we want our documents to be kept in the system. Documents that have exceeded the retention period will be automatically expunged. This could be important for regulatory compliance or for managing the overall storage size.

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

3.1.2 Enrich Tab

_1. Click on the ENRICH tab.

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



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

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

3.1.3 Configure Tab

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

The screenshot shows the 'Configure' tab for the 'Clandis Baker Project'. In the 'Import / Export ontology' section, there are three buttons: 'Export project' (highlighted in blue), 'Language settings', and 'Git server configuration'. Below these, under 'Import project', there is a note stating 'You cannot import a project if the current project has been deployed.' and a 'Import project' button.

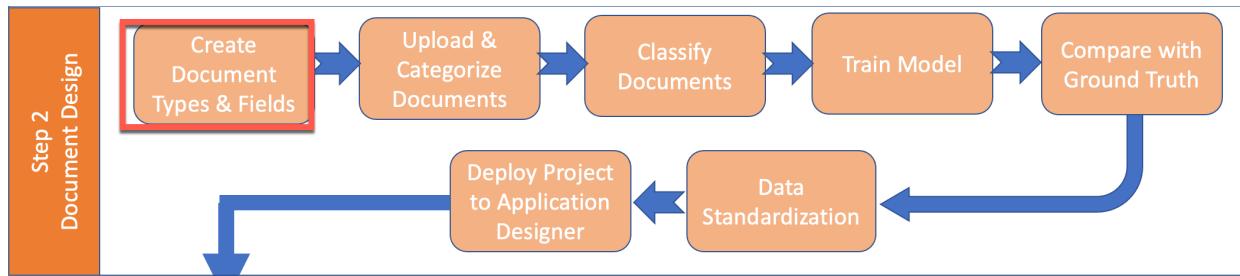
In Extraction language, select which languages are used in the documents that you plan to process. You can choose English, Dutch, French, German, Brazilian Portuguese, or Spanish. Make sure to choose only the language or languages that are likely to be used in your document sets. Choosing more than one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

The screenshot shows the 'Configure' tab for the 'Clandis Baker Project'. Under 'Language settings', the 'Extraction language' section is active. It includes a note about choosing extraction languages, a 'Default' section with 'English' checked, and a 'Other languages' section with checkboxes for Dutch, French, German, Portuguese (Brazil), Spanish, and French. The 'Display name language' section is also shown, with a dropdown set to 'English (en) (default)'. At the bottom right are 'Cancel' and 'Save' buttons.

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

1.9 Configure a Wage and Tax document type.



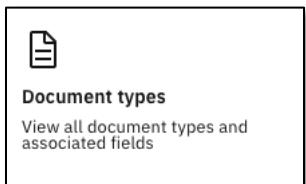
Before we use the guided configuration, you will configure some additional document types and fields used to extract data prior to uploading sample documents.

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

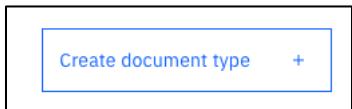
1.10 Create Wage and Tax document type.

- _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 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.

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

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

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.	
<input type="checkbox"/> 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.

_14. Click the ADD button.

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

_15. Select your Wage and Tax type. On the right, you should see an empty table of fields.

1.11 Create Field

We can now add some fields to the class.

_1. Click ADD FIELDS

_2. Enter the following values under the GENERAL Settings header

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

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

You should now see the following:

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

_3. Click the NEXT button.

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

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

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

Name	Type	Required	Sensitive
Federal Income Tax Withheld	Decimal	true	false

1.12 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

_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.

1.13 Create Employee Social Security Number Field

_1. Click on ADD FIELDS



Enter the following values in the GENERAL page.

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

Your screen should now look like the image below:

The screenshot shows the 'Employee Social Security Number' field configuration screen. At the top, there are tabs for 'General' (selected), 'Field patterns', and 'Value settings'. The 'General' tab has fields for 'Display name' (Employee Social Security Number) and 'Symbolic name' (EmployeeSocialSecurityNumber). It also includes checkboxes for 'This field is required' and 'This field contains sensitive information'. The 'Value settings' tab shows an 'Aliases' section with a text input field containing several suggestions: 'a Employee's social security number', 'a Employee's social security no.', 'a Employee's SSA number', 'Employee social security number', and 'a Employee Social Security Number'. There are 'Cancel' and 'Next' buttons at the top 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

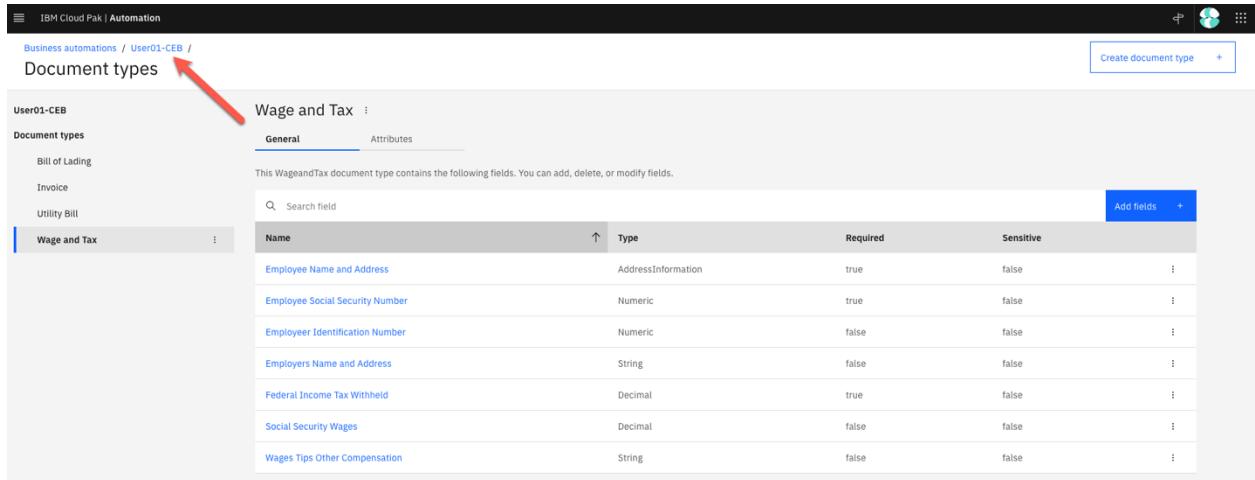
Field Name	Description	Type	Mandatory	Aliases
Employer Identification Number		sys:String	N	<ul style="list-style-type: none"> • b Employer identification number (EIN) • b Employer's FED ID number • b. Employer ID number • Employer identification number
Employers Name and Address		sys:String	N	<ul style="list-style-type: 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>.

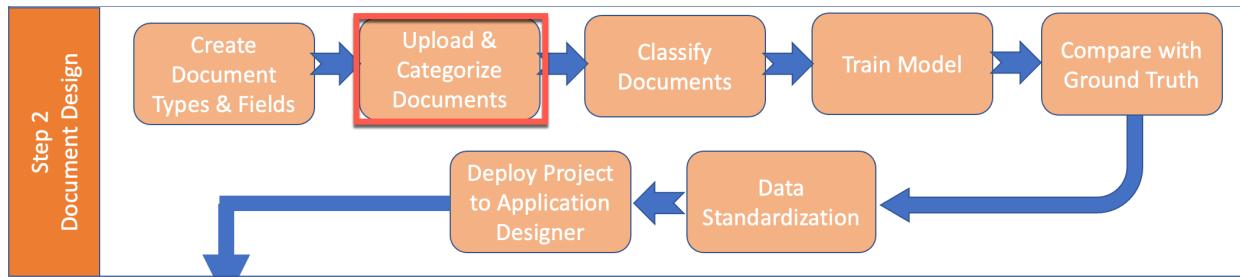


The screenshot shows the 'Document types' page within the 'User01-CEB' project. The breadcrumb navigation bar at the top displays 'IBM Cloud Pak | Automation', 'Business automations / User01-CEB / Document types'. The main content area is titled 'Wage and Tax' and includes tabs for 'General' and 'Attributes'. Below the tabs, a note states: 'This WageandTax document type contains the following fields. You can add, delete, or modify fields.' A search bar labeled 'Search field' is present. A table lists document fields with columns for 'Name', 'Type', 'Required', and 'Sensitive'. The table rows include:

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

A blue button labeled 'Add fields +' is located in the top right corner of the table area. The sidebar on the left lists other document types: 'Bill of Lading', 'Invoice', 'Utility Bill', and 'Wage and Tax' (which is selected).

4 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

1.14 Categorize documents.

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

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

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

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

Let's see what that looks like.

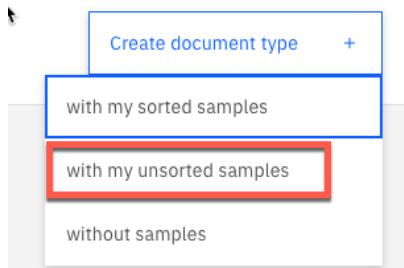
_2. Click on **CREATE DOCUMENT TYPE** in the top right of the screen.

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

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

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

_3. Select the second option titled with my unsorted samples



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

_4. Click Upload to get document samples.

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

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

_5. Click on the CATEGORIZE button.

The screenshot shows a web-based interface for document classification. At the top, there's a navigation bar with 'IBM Cloud Pak | Automation' and a search bar. Below that, a breadcrumb trail shows 'Business automations / User01-CEB / Document types and samples / Create document types'. On the right, there are 'Cancel' and 'Categorize' buttons. A sidebar on the left has 'Upload unsorted documents' and 'Review categories' options. The main area is titled '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.' It features a search bar 'Search sample documents' and an 'Upload' button with a file icon. A list of 12 PDF files is shown in a table format:

Document name
Mortgage Agreement1.pdf
Mortgage Agreement2.pdf
Mortgage Agreement3.pdf
Mortgage Agreement4.pdf
Mortgage Agreement5.pdf
TR_FW2_1001_0000_PS.pdf
TR_FW2_2000_0000_PS.pdf
TR_FW2_3000_0000_PS.pdf
TR_FW2_3001_0000_PS.pdf
TR_FW2_4000_0000_PS.pdf
UBILLCable_081_1_11.pdf
UBILLCable_082_1_11.pdf

At the bottom, there are dropdown menus for 'Items per page' (set to 20), a page number indicator '1 of 1 page', and navigation arrows.

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

You will need to:

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

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

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

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.

IBM Cloud Pak | Automation

Business automations / User01-CEB / Document types and samples / Create document types

Review each category, verify the documents, and assign each category to a new or pre-trained document type. Learn more

Categories (3)

- Category 1
- Category 2
- Category 3

Document types (4)

- Bill of Lading (21 samples)
- Invoice (31 samples)
- Utility Bill (25 samples)
- Wage and Tax (0 samples)

Category 1 sample documents (2)

Search sample documents

Upload ↑

- Document name
- UBILLCable_081_1_1.pdf
- UBILLCable_082_1_1.pdf

IBM Cloud Pak | Automation

Business automations / User01-CEB / Document types and samples / Create document types

Review each category, verify the documents, and assign each category to a new or pre-trained document type. Learn more

Categories (3)

- Category 1
- Category 2
- Category 3

Document types (4)

- Bill of Lading (21 samples)
- Invoice (31 samples)
- Utility Bill (25 samples)
- Wage and Tax (0 samples)

Category 2 sample documents (5)

Search sample documents

Upload ↑

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

IBM Cloud Pak | Automation

Business automations / User01-CEB / Document types and samples / Create document types

Review each category, verify the documents, and assign each category to a new or pre-trained document type. Learn more

Categories (3)

- Category 1
- Category 2
- Category 3

Document types (4)

- Bill of Lading (21 samples)
- Invoice (31 samples)
- Utility Bill (25 samples)
- Wage and Tax (0 samples)

Category 3 sample documents (5)

Search sample documents

Upload ↑

- Document name
- TR_FW2_1001_0000_P5.pdf
- TR_FW2_2000_0000_P5.pdf
- TR_FW2_3000_0000_P5.pdf
- TR_FW2_3001_0000_P5.pdf
- 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.

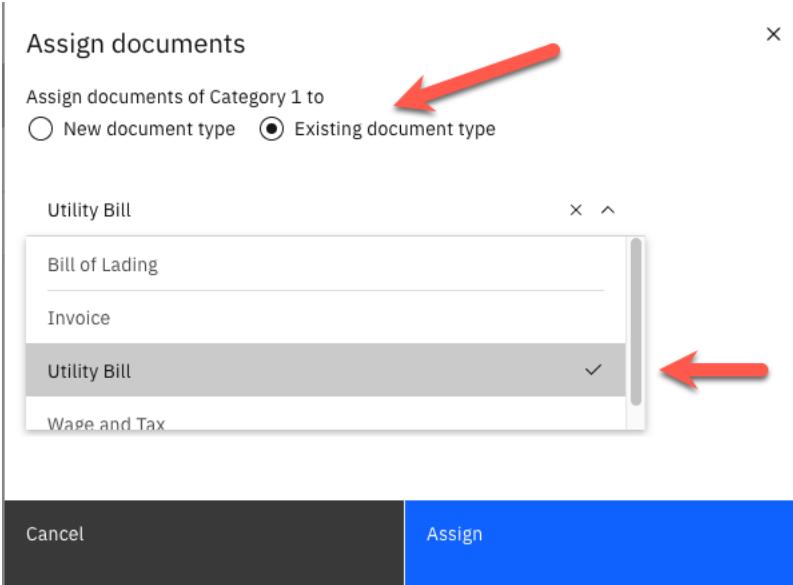
_7. If all documents within a category are correct as illustrated in the following screen shot, hover over the category name and **Click on the 3 dots** at the end of the category name.

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

_8. Select ASSIGN TO DOCUMENT TYPE

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

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



_10. Click Assign to close the dialog box

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

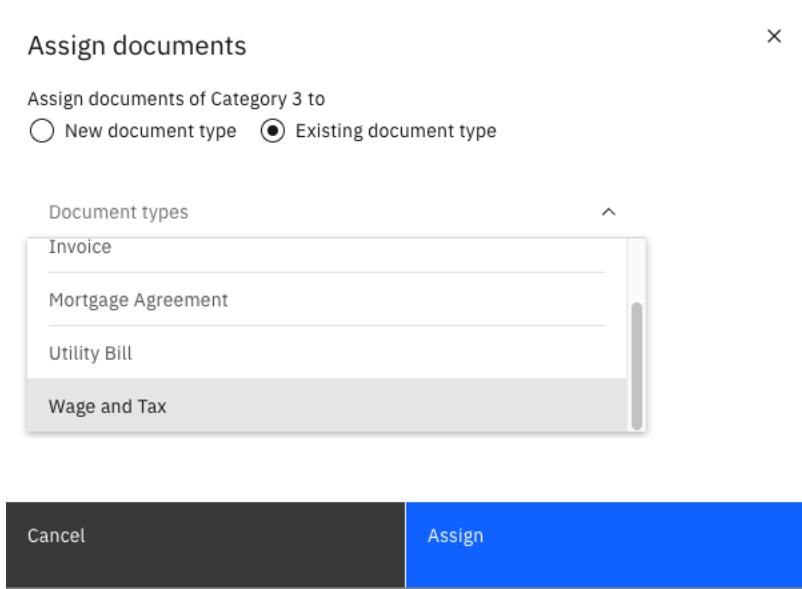
_11. Select the next Category 2 and Click on the 3 dots and Select Assign these documents to a document class.

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

_13. Enter Mortgage Agreement in the field

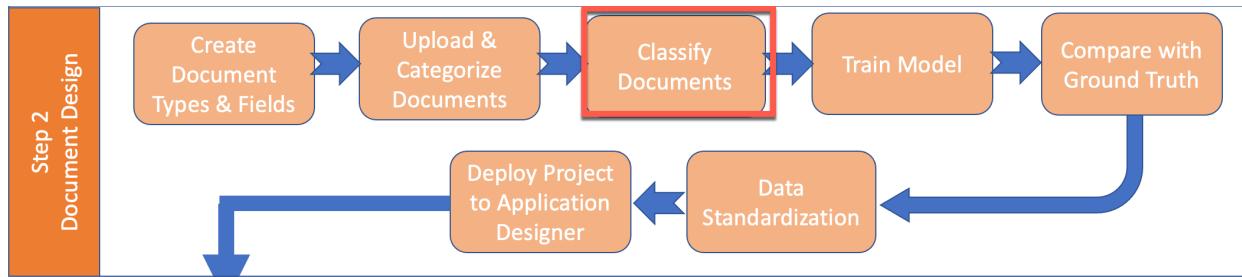
The screenshot shows the 'Assign documents' dialog box. At the top, it says 'Assign documents of Category 2 to'. Below that are two radio buttons: 'New document type' (checked) and 'Existing document type' (unchecked). The main area has a 'Document type display name' field containing 'Mortgage Agreement'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

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



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

5 Train classification



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

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

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

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

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

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

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

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

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

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

Classification model

Accuracy 84.8%

Document types

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

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

Mortgage Agreement sample documents (5) Training/test ratio in % 100/0 Auto generate 70/30 split

Training set (5)	Test set (0)
100% of total samples	0% of total samples
Search training set sample documents	Search test set sample documents
Mortgage Agreement1.pdf	
Mortgage Agreement2.pdf	
Mortgage Agreement3.pdf	
Mortgage Agreement4.pdf	
Mortgage Agreement5.pdf	

There are no documents in the test set. Include at least 1 document in the test set to view training results.

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

Classification model

Accuracy 84.8%

Document types

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

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

Mortgage Agreement sample documents (5) Training/test ratio in % 60/40 Auto generate 70/30 split

Training set (3)	Test set (2)
60% of total samples	40% of total samples
Search training set sample documents	Search test set sample documents
Mortgage Agreement3.pdf	Mortgage Agreement1.pdf
Mortgage Agreement4.pdf	Mortgage Agreement2.pdf
Mortgage Agreement5.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.

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

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

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

You should see something like the following:

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

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

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

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

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

_9. Click Done

1.15 How do I improve my results?

Option 1 – Add more samples.

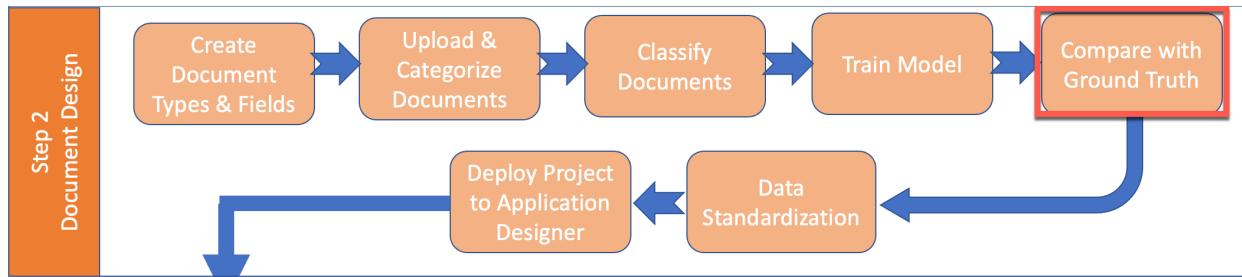
Option 2 – review all uploaded samples.

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

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

.

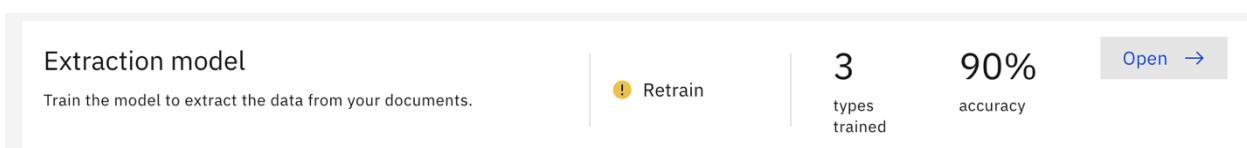
6 Data extraction



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

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

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



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

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

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

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



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

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

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

_5. Click on **Teach Samples**.

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

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

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

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

Field Name	Value Captured
Federal Income Tax Withheld	abc Text Required <input type="checkbox"/> Field label (optional) <input type="checkbox"/> Draw Captured field label <input type="checkbox"/> Field value <input type="checkbox"/> Draw Captured field value
Employee Name and Address	abc Text Required
Employee Social Security Number	abc Text Required
Employer Identification Number	abc Text
Employers Name and Address	abc Text
Social Security Wages	abc Text

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.

1.16 Correcting extracted values

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

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

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

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

Notice a green check mark signifies this field is complete.

The screenshot shows the IBM Cloud Pak Administration interface with the W-2 Wage and Tax Statement form loaded. The "Save match" operation has been completed, as indicated by the green checkmark next to the "Saved!" message in the pop-up window. The "Federal Income Tax Withheld" field now contains the value "123456789.99".

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

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

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

- Document:** TR_FW2_1001_0000_PS.pdf
- Fields Extracted:**
 - a Employee's social security number:** 577-22-3048
 - b Employer identification number (EIN):** 14-023285
 - c Employer's name, address, and ZIP code:** Long Lengthy Name The Corporation, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234
 - d Control number:** 123456 A78
 - e Employee's name and initial:** Michael R. Scott, David S.
 - f Employee's address and ZIP code:** MIN 123456789
 - g State:** MN
 - h State income tax number:** 123456789.99
 - i State wages, tips, etc.:** 123456789.99
 - j State income tax:** 123456789.99
 - k Local wages, tips, etc.:** 123456789.99
 - l Local income tax:** 123456789.99
 - m Local tax name:** ABCDEFGHI
- Field Labels and Values Captured:**
 - Federal Income Tax Withheld:** 123456789.99
 - Employee Name and Address:** Michael R. Scott, David S.
 - Employee Social Security Number:** 123456789.99
 - Employer Identification Number:** 123456789.99
- Actions:**
 - Draw:** Buttons for drawing field values.
 - Save selection:** Button to save the selected field values.
 - Mark this document as ready for training:** Checkbox for marking the document as ready for training.

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

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

The screenshot shows the IBM Cloud Pak Administration interface. On the left, a PDF of a W-2 form is displayed. The form includes fields for Employee's social security number (577-22-3048), Employer identification number (EIN) (14-023285), Employer's name, address, and ZIP code (Long Lengthy Name The Corporation, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234), Control number (123456 A78), Employee's first name and initial (Michael Robert David Smithson III), Employee's address and ZIP code (MN, 123456789), and various wage and tax amounts. On the right, the extracted fields are listed in a table:

Field Name	Value Captured
Federal Income Tax Withheld	abc 123456789.99 Required
Employee Name and Address	Michael Robert David Smithson III 56334 Full Sized Avenue Unit 1234 Minneapolis, Minnesota 55411-1234
Employee Social Security Number	Text
Employer Identification Number	Text
Mark this document as ready for training.	<input type="checkbox"/>

Below the table, there are buttons for 'Draw' and 'Capture subfields'. At the bottom, there are 'Previous sample' and 'Next sample' buttons, and a 'Save selection' button.

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

The screenshot shows the IBM Cloud Pak Administration interface. On the left, a PDF of a W-2 form is displayed. The form includes fields for employer information (Employer identification number: 14-023285, Employer name and address: Long Lengthy Name The Corporation, 56334 Full Sized Avenue Unit 1234, Minneapolis, Minnesota 55411-1234), employee information (Employee's social security number: 577-22-3048, Employee's first name and initial: Michael Robert David Smithson III, Employee's address and ZIP code: MN 123456789), and tax details (State wages, tips, etc.: 123456789.99, State income tax: 123456789.99, Local wages, tips, etc.: 123456789.99, Local income tax: 123456789.99). The year is 2020. On the right, a list of extracted fields is shown with their captured values. A red arrow points from the 'Field Value' section of the 'Wages, tips, other compensation' entry to the 'Mark this document as ready for training' checkbox at the bottom of the interface.

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

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

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

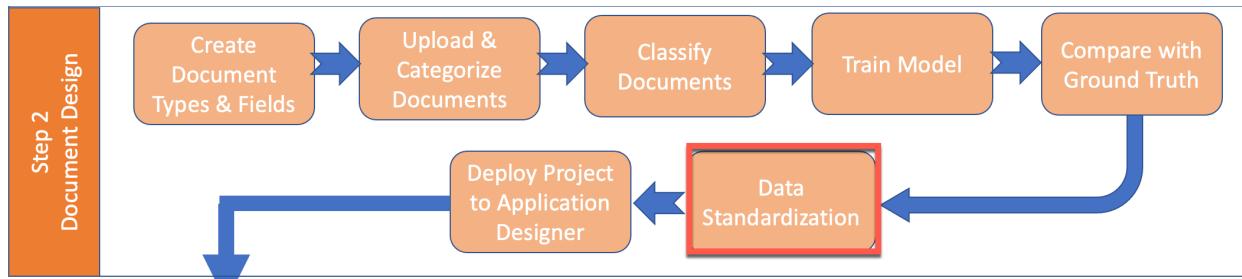
The screenshot shows the IBM Cloud Pak Administration interface. On the left, there are two W-2 form documents. The top one is titled "2020 W-2 and EARNINGS SUMMARY". It contains various fields such as Employee Name (Michael Robert David Smithson III), Social Security Number (123456789.99), and wages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

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

1.17 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.

7 Data standardization

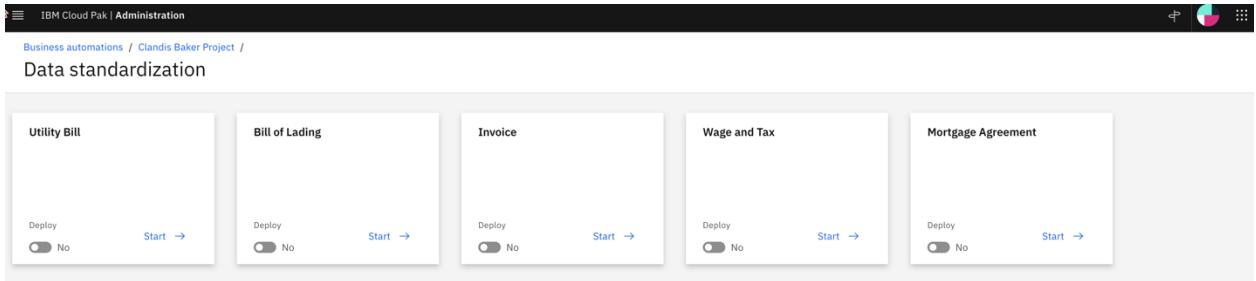


Next, we need to standardize the data that will be presented in the user interface and how it will be stored in the FileNet repository for example. Data standardization is the process of defining attributes for a data field in a standardized way. This is done using data definitions. These definitions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

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

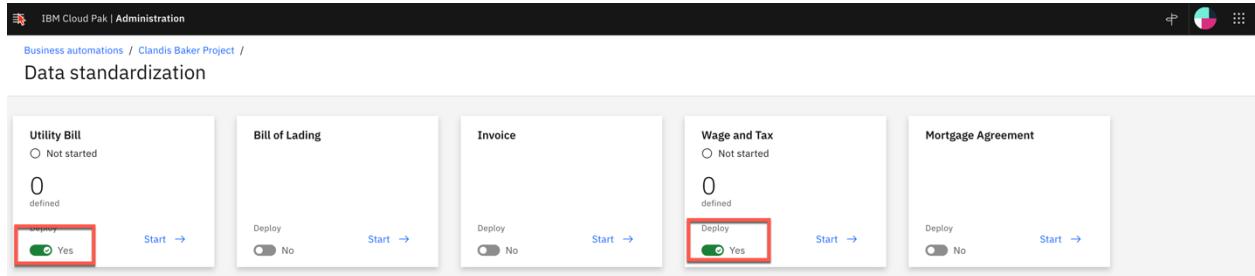
Model Type	Status	Count	Accuracy / Samples
Document types and samples	Ready	4 types	22 samples on average
Classification model	Ready	3 types trained	100% accuracy
Extraction model	Ready	3 types trained	97% accuracy
Data standardization	Not ready		

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



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

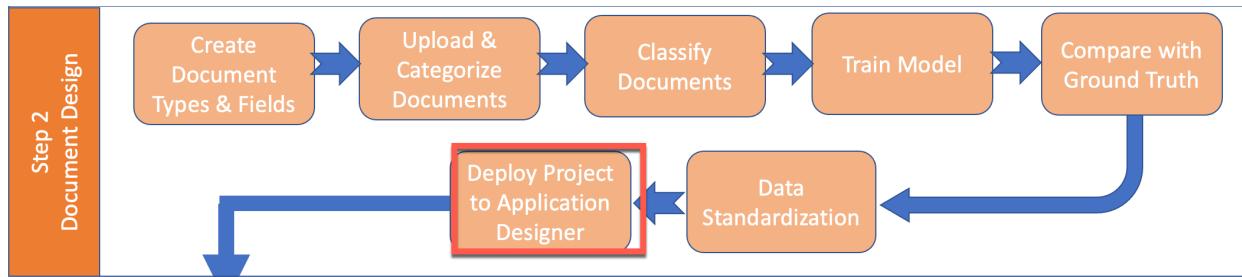
_3. Click on Start



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

Business automations / Clandis Baker Project /

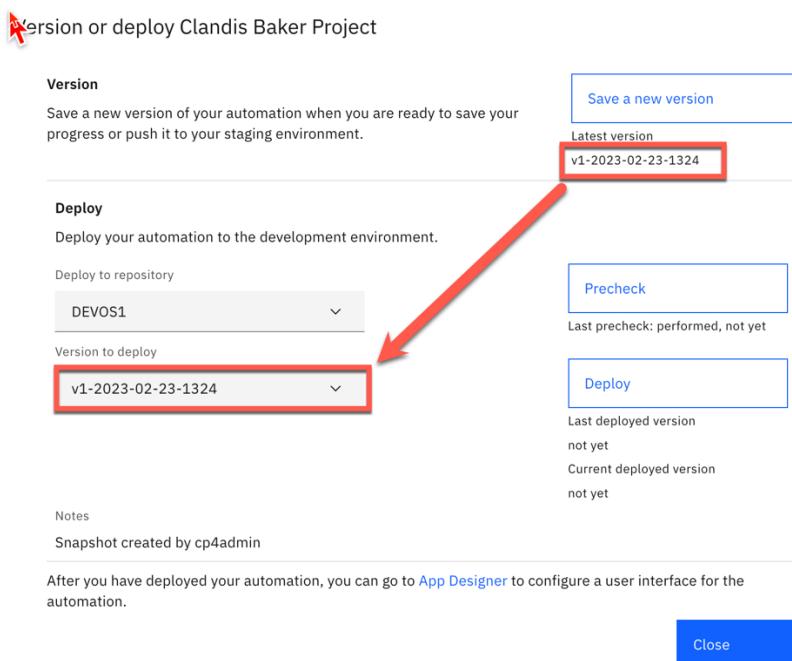
8 Version and deploy your project



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

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

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



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

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

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

The screenshot shows the 'Version or deploy Clandis Baker Project' dialog. The 'Deploy' section is highlighted with a red box. Inside this box, the 'Deploy' button is shown with the text 'Last deployed version v1-2023-02-23-1324' and 'Current deployed version v1-2023-02-23-1324' below it.

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

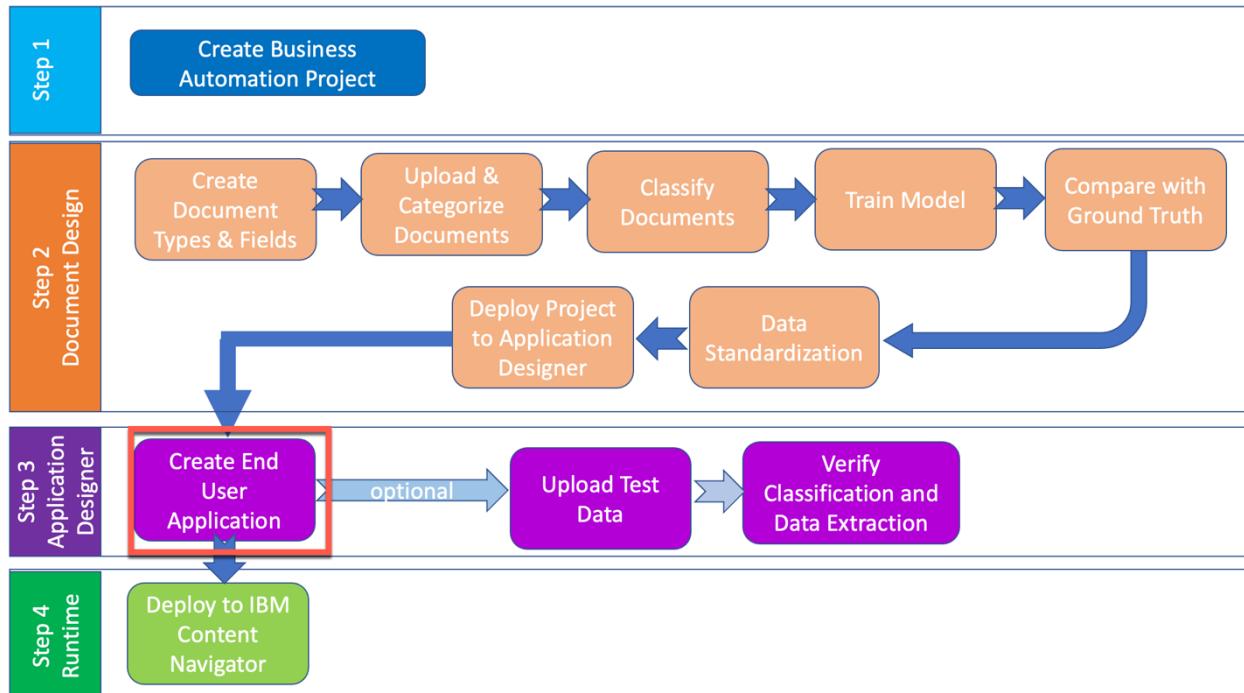
5. Click **Close** button.

Once deployed, proceed to the next steps.

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

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

9 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

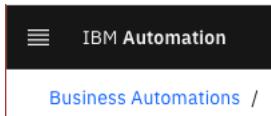
1. Batch Document Processing template – used to process batches of documents.
2. Document Processing Template – used to process single documents.

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.

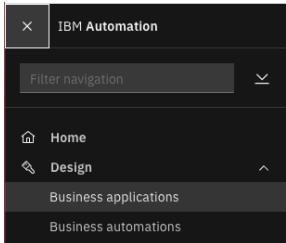
Changes to the application itself will not be in the scope of this lab.

1.18 Create your Runtime Application.

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



and **selecting Business Applications**



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

Business applications

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

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

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

Create a business application

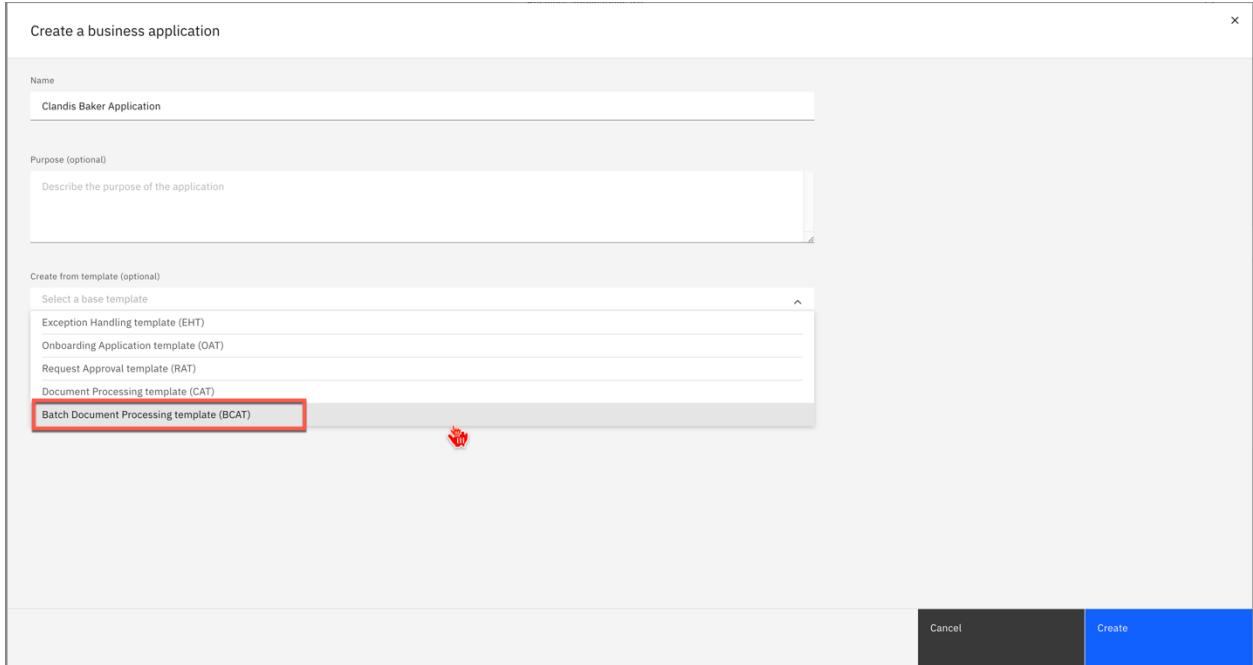
Name

Purpose (optional)
Describe the purpose of the application

Create from template (optional)
Select a base template

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

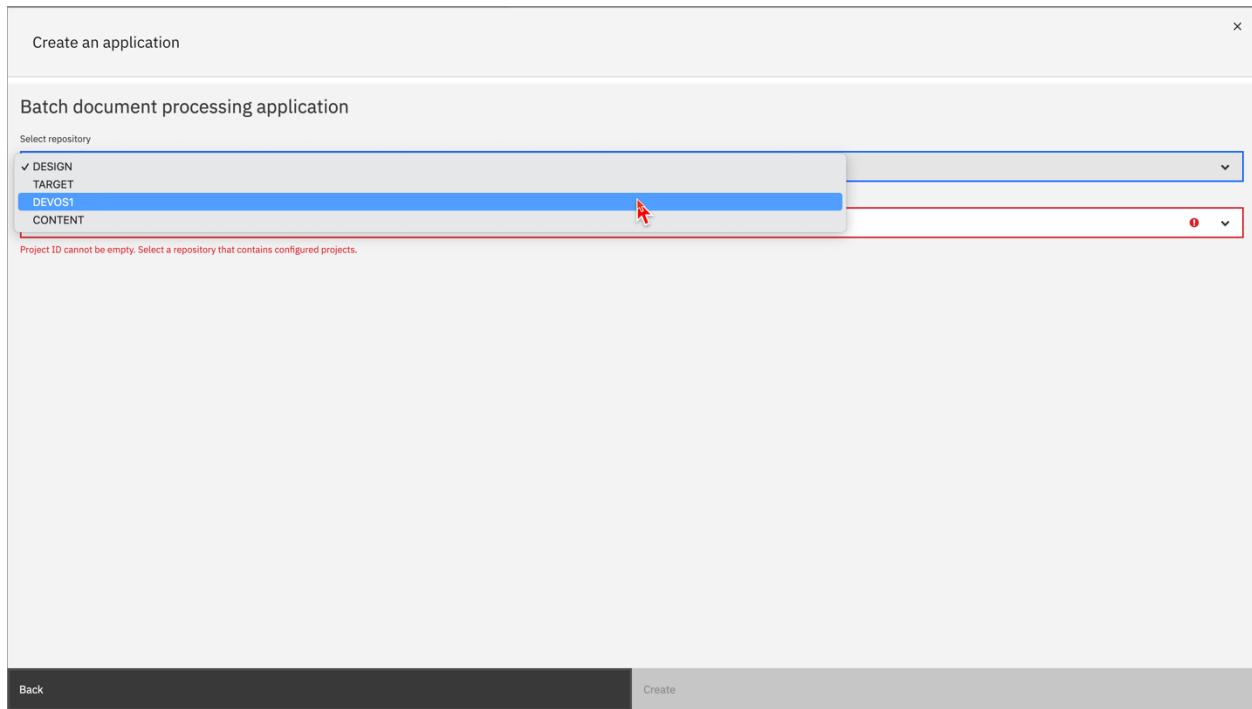
- _4. In the Create Form Template drop down **select Batch Document Processing template (BCAT)**.



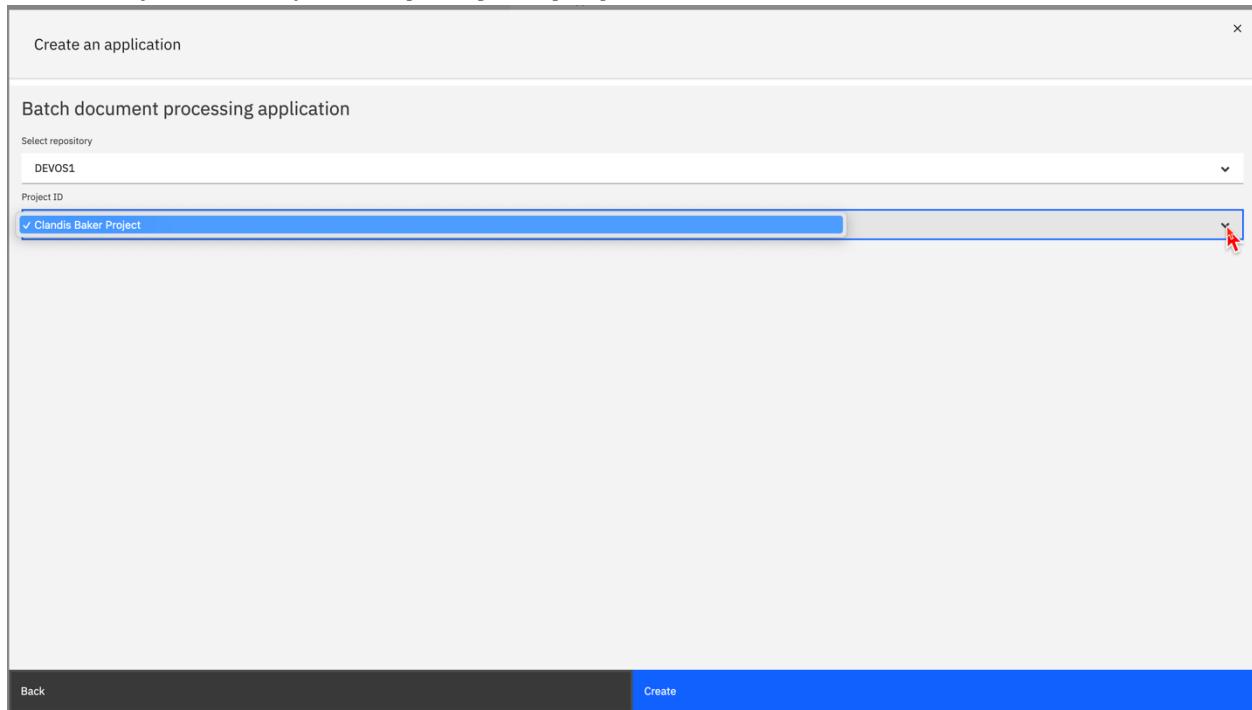
You could have selected the Document Processing Template if you only wanted to process a single document at a time, but in this lab, you will process several documents in a batch.

- _5. Click Next**

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



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



_8. Click **Create**

You should now be in the *Application Designer*

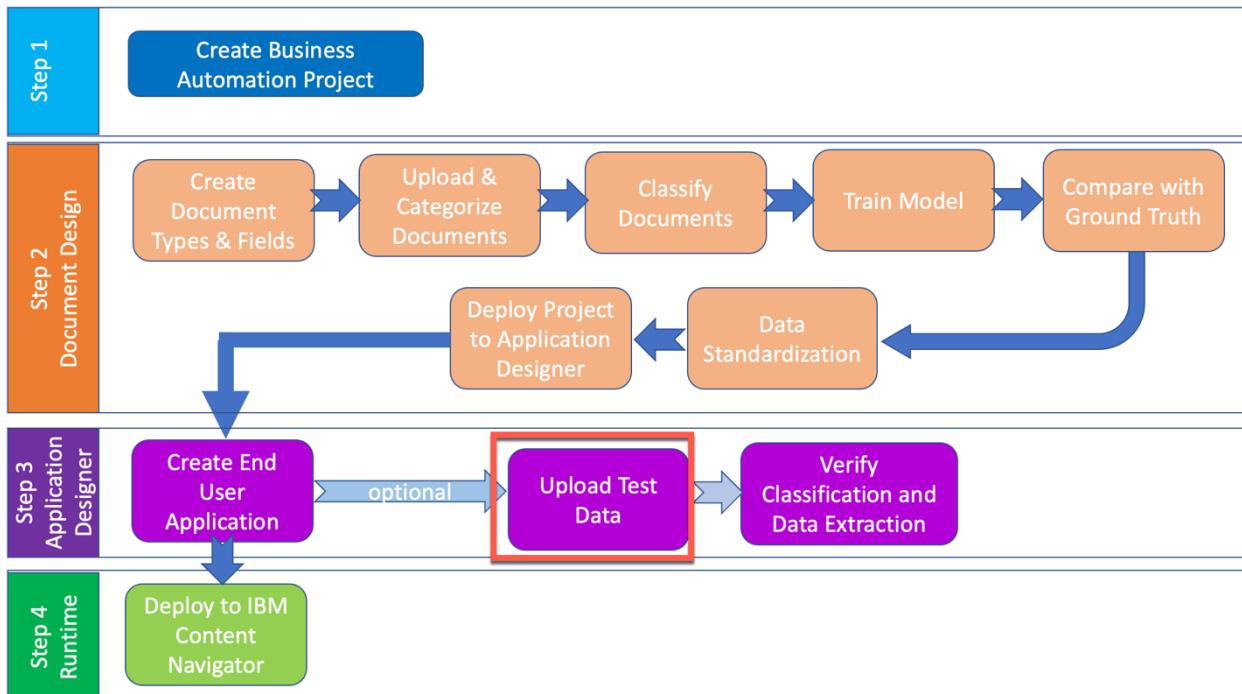
The screenshot shows the IBM Cloud Pak Application Designer interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Administration' and a 'Preview' button. Below the navigation, the application title 'Clandis Baker Application' is displayed. The main content area shows 'Review batch issues' with two tabs: 'Document type and page order issues' and 'Data extraction issues'. Below this, a 'Batches' section shows a 'Content List' with three items: 'My Document1', 'My Document2', and 'My Document3'. Each item has columns for Name, Size, Modified by, Last modified, and Version. A blue 'Add' button is visible at the top of the list. On the right side, there's a sidebar titled 'Drag a component to your page' containing a grid of UI component icons, such as 'Add batch modal', 'Add document modal', 'Add folder modal', 'Batch content', 'Button', 'Check box', 'Collapse panel', 'Content list', 'Content properties', 'Data verify...', 'Date/time picker', 'Delete object modal', 'Display text', 'Document correction', 'Document reference', 'Document thumbnail', 'Document viewer', 'Edit properties', and 'Export document'.

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

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

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

1.19 Upload documents for processing



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

IBM Cloud Pak | Administration x Clandis Baker Application x +

Not Secure https://cpd-cp4ba-starter.apps.ocpinstall.gym.lan/ae-pbk/Clandis%20Baker%20Application(CBA)?locale=en

Bookmark this w3 Doc Imaging and... CoC - Home - Dat... Customer Log In ... Sign In - Skytap WW SWAT Seismic - Login ECM Enablement... Advanced Case M... ADP IBM GitHub Login - Jazz Team... Other Bookmarks

Review batch issues

Document type and page order issues 0 batches

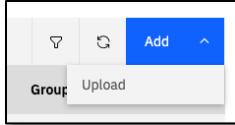
Data extraction issues 0 batches

Batches

Name	Files	Priority	Status	Added on	Added by	Group	Location	Add
No items found.								

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

_2. Click on Add, then Upload.



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

Upload new batch

* Display Name
Batch 1

Description

Priority
High

- _4. Click **Select files**.

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

- _5. Click **Open**

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

Add Files

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

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

Cancel Add

_6. Click on the Add button.

The screenshot shows the 'Review batch issues' section with two tiles: 'Document type and page order issues' (0 batches) and 'Data extraction issues' (0 batches). Below this is a table titled 'Batches' with columns: Name, Files, Priority, Status, Added on, Added by, Group, and Location. A single row is shown for 'Batch01'. At the bottom left is a pagination bar 'Items per page: 100 1-1 of 1 items'. At the top right is a blue 'Add' button with a dropdown arrow.

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

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

The screenshot is similar to the previous one, but the status column for 'Batch01' now shows a blue progress bar with the text '3 of 5 files processed'. A red arrow points to the three-dot menu icon (three vertical dots) at the end of the status bar for the first batch row.

_8. Click Submit

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

The screenshot shows the 'Review batch issues' section with the 'Document type and page order issues' tile now displaying '1 batches'. The 'Data extraction issues' tile remains at 0 batches. Below is the 'Batches' table, which now shows a single row for 'Batch 1'. A red box highlights the 'Document type and page order issues' tile.

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

Batch Document Processing Application /

Document type and page order issues

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

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

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

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

Batch01

Cancel | Save changes | Submit

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

Edit actions for .PDF and .TIFF



1

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

Employee's social security number 14-023265	Employer identification number 577-22-3048	State No. 1040-0008
Employee's name, address, and ZIP code Terry L. Stacey 563 Stoney Brook Rd Minneapolis, MN 55411	Wages, tips, other compensation 18000.00	Federal income tax withheld 1800.00
Tax withheld 17700.00	Social security tax withheld 1113.33	State income tax withheld 261.00
Medicare wages and tips 18000.00	Medicare tax withheld 261.00	Alcohol excise tax 400.00
Social security tax withheld 400.00	Health care benefit 443.21	
Control number 210220 A13	Nonqualified plan 300.00	State A 256.00
Employee's name and ZIP code Benjamin P. Charles 4326 Aldrich Rd Minneapolis, MN 55412	Plan B 20000.00	Local C 532.00
State M 1280.00	Test form D 423.00	Local M 500.00
Local M 17700.00		State M 500.00

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

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

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

Batch01

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

Select document type

Document type: Wage and Tax (Recommended)

Cancel Update

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

Detailed description of the form fields (e.g., Name, Address, Social Security Number, etc.)

Submit Save changes

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



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

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

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

[Batch Document Processing Application](#) /
Document type and page order issues

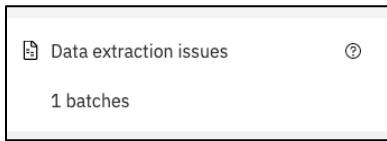
1.21 Correct extraction issues

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

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

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

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



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



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

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

Items per page: 100 1-5 of 5 items

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

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

The screenshot shows a document processing interface with a W-2 form on the left and its extracted data on the right.

W-2 Form Data:

Employee's social security number	577-22-3048	OMB No. 1445-0008			
Employer identification number (EIN)	14-023285	1. Wages, tips, other compensation	18000.00	4. Federal income tax withheld	\$1800.00
Employer's name, address, and ZIP code	Test and Rest Inc. 563 Stoney Brook Rd Minneapolis, MN 55411	2. Social security wages	17700.00	4. Social security tax withheld	1113.33
Control number	210220 A13	3. Medicare wages and tips	18000.00	5. Medicare tax withheld	261.00
Employee's first name and initial	Last name	6. State nonqualified plans	300.00	7. Social security tips	400.00
Bernard P. Aldrich	Aldrich	7. State income tax	1260.00	8. Allocated tips	400.00
4326 Aldrich Rd		8. State wages, tips, etc.	17700.00	9. Dependent care benefits	543.21
Minneapolis, MN 55412		10. Local wages, tips, etc.	500.00		
Employee's address and ZIP code	11. Local income tax	12. Local wages, tips, etc.	500.00		
12. Local income tax	13. Local wages, tips, etc.	14. Other	MPLS		
13. Local income tax	14. Other				
14. Other					
Form W-2 Wage and Tax Statement 2020 Department of the Treasury - Internal Revenue Service					

Extracted Data:

- Federal Income Tax Withheld ***
1800.00
- Social Security Wages**
17700.00
- Wages Tips Other Compensation**
18000.00
- Employee Social Security Number ***
577-22-3048
- Employer Identification Number ***
14-023285
- Employee Name and Address ***
 - Organization**
(none)
 - Name**
(none)

This is reason for seeing the Data Issues in the previous screen. Looks like we'll need to add more documents for the Wage and Tax!

Take a moment to discover the image viewer features at top:

The screenshot shows the same document processing interface after adding validation errors to the extracted data.

Fields with issues:

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

- Rotate image
- Visual effect adjustment

- Invert

Image viewer features at bottom:

The screenshot shows a document processing application with the following components:

- Top Bar:** TR_FW2_1001_0001_PS.pdf | Document type: Wage and Tax
- Left Panel:** A thumbnail view of the W-2 form.
- Middle Panel:** The full W-2 form (Form 1099) for Employee Social Security Number 577-22-3048, Employer Identification Number 14-023285, and Control Number 210220 A13. The form details wages, taxes withheld, and other financial information for the year 2020.
- Right Panel:** An "Extracted data" section showing validation errors for the Employee Social Security Number and Employer Identification Number fields.
- Bottom Panel:** A toolbar with icons for zoom, search, and navigation, along with a page number indicator (1 / 1).

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

Field features

This screenshot is identical to the one above, but the "Fields with issues" section in the right panel is highlighted with a red box. This section lists validation errors for the Employee Social Security Number and Employer Identification Number fields.

- Show all fields.
- Show fields with issues.

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

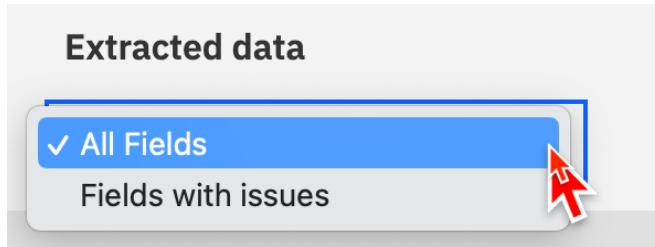
The screenshot shows the 'Extracted data' view for a W-2 form. On the left, the W-2 form is displayed with various fields filled in. On the right, the 'Extracted data' panel shows a list of fields with their values. Two fields are highlighted with red boxes and have red validation error icons next to them:

- Employee Social Security Number ***: Value: 577-22-3048
- Employer Identification Number**: Value: 14-023285

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

- Validation error: Employee Social Security Number *
- Validation error: Employer Identification Number

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



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

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

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

The screenshot shows the ADP software interface. On the left is the PDF document of the W-2 form. On the right is the 'Extracted data' panel. The panel has tabs for 'All Fields' and 'Fields with issues'. Under 'All Fields', several fields are listed with their extracted values. Under 'Fields with issues', there are no items listed.

Field	Value
Federal Income Tax Withheld	1800.00
Employer Name and Address	Test and Rest Inc. 563 Stoney Brook Rd Minneapolis, MN 55411
Social Security Wages	17700.00
Wages, Tips, Other Compensation	18000.00
Social Security Tax Withheld	1113.33
Medicare Wages and Tips	1800.00
Medicare Tax Withheld	261.00
Social Security Tips	400.00
Allocated Tips	400.00
Dependent Care Benefits	543.21
State Income Tax Withheld	300.00
State Wages, Etc. Withheld	18000.00
Local Income Tax Withheld	17700.00
Local Wages, Etc. Withheld	500.00
Local Income Tax Withheld	MPLS
State Wages, Etc. Withheld	18000.00
Local Income Tax Withheld	1260.00
Local Wages, Etc. Withheld	17700.00
Local Income Tax Withheld	500.00
Local Wages, Etc. Withheld	MPLS

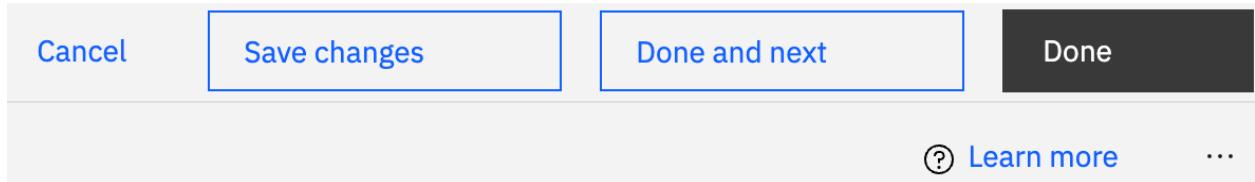
If we change the Extracted data back to Fields with issues:

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

The screenshot shows the ADP software interface. The 'Fields with issues' tab is selected in the 'Extracted data' panel. A message 'There aren't any extraction issues' is displayed. The rest of the interface is identical to the previous screenshot.

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

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



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

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

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

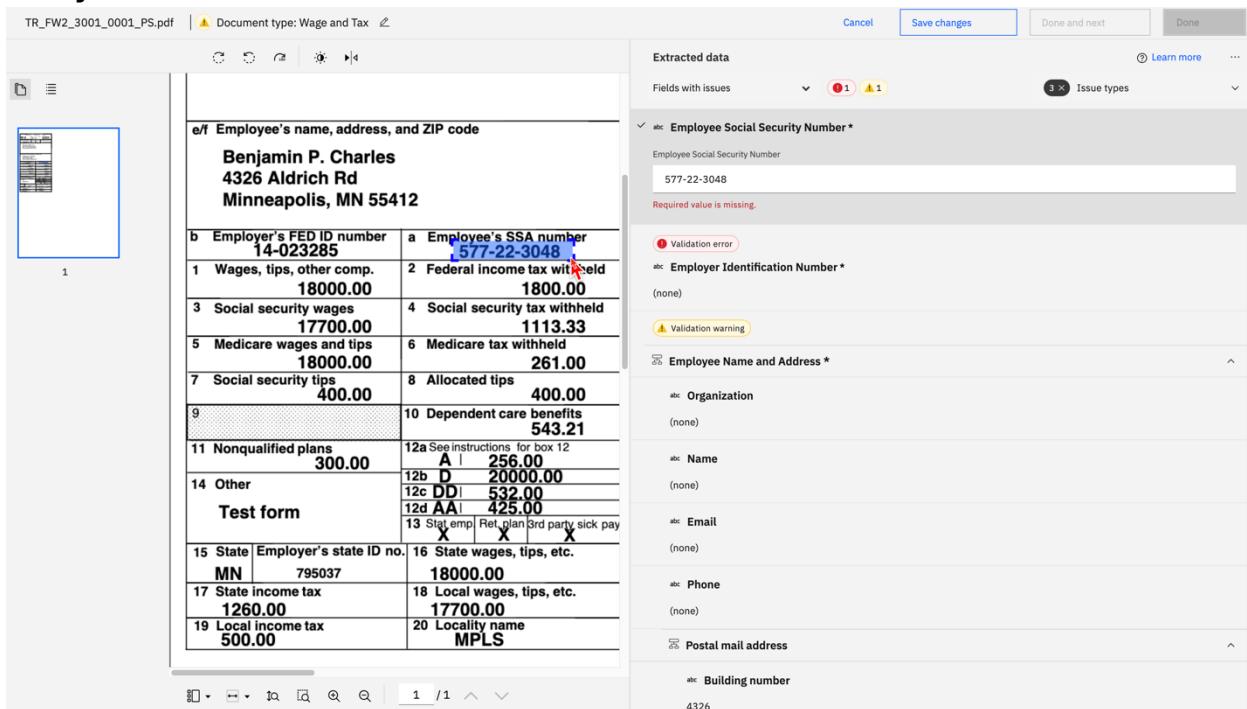
The screenshot shows a document viewer window with a PDF of a W-2 form. To the right is an 'Extracted data' panel. The panel has a header with buttons for 'Cancel', 'Save changes', 'Done and next', and 'Done'. Below the header is a 'Learn more' link and a '...' button. The main area of the panel is titled 'Extracted data' and contains a 'Fields with issues' dropdown. Underneath it, there are three sections with validation status indicators (red circle for error, yellow triangle for warning):

- Employee Social Security Number ***: Validation error (none)
- Employer Identification Number ***: Validation error (none)
- Employee Name and Address ***: Validation warning (none)

Below these sections are expandable sections for 'Organization', 'Name', 'Email', 'Phone', 'Postal mail address', and 'Building number'.

_10. Click on the Employee Social Security Number.

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

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


The screenshot shows a document processing application interface. On the left is a preview of the PDF document, which contains a tax form with various fields filled in. One specific field, "Employee's SSA number" (containing "577-22-3048"), is highlighted with a blue rectangular selection box. To the right of the PDF is a panel titled "Extracted data" showing the corresponding data fields and their values. The "Employee Social Security Number" field is listed with the value "577-22-3048" and a note below it stating "Required value is missing." There are also other fields listed such as "Employer Identification Number" and "Employee Name and Address".

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

END OF LABS

10 Export Import Project.

From the Business Automations

1. From the Business Automations screen select Document Processing.

The screenshot shows the 'Business automations' screen in the IBM Cloud Pak interface. At the top, there's a navigation bar with 'IBM Cloud Pak | Administration'. Below it, a sidebar on the left lists categories: 'Published automation services', 'Decision' (which is expanded), 'Workflow', and 'External'. Under 'Decision', 'Document processing' is selected and highlighted with a blue border. On the right, a main panel displays 'Document processing automations (1)' with a single entry: 'Clandis Baker Project' last edited on 02/23/2023. A 'Create' button is at the bottom left, and a 'Import' button is at the bottom right.

2. Select <your project name> Click open

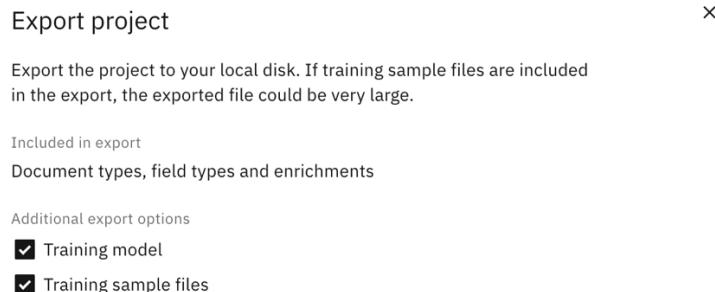
This screenshot is similar to the previous one, showing the 'Business automations' screen. The 'Document processing' project is selected. A red arrow points to the 'Open' button, which is highlighted with a red border. The rest of the interface elements are identical to the first screenshot.

3. From the Main screen select the Configure tab

This screenshot shows the configuration screen for the 'Clandis Baker Project'. The top navigation bar has 'IBM Cloud Pak | Administration' and the project name 'Clandis Baker Project'. Below it, there are tabs for 'Build', 'Enrich', and 'Configure', with 'Configure' being the active tab. A red arrow points to the 'Import / Export ontology' section on the left. On the right, there are sections for 'Share' (last shared 2 hours ago) and 'Version / Deploy' (latest version v2 deployed 2 hours ago). The central area contains sections for 'Export project' (button 'Export project') and 'Import project' (button 'Import project').

_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

10.1 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>

10.2 Can't find user/password in Daffy

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

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

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

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

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

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

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

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

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

Environment verification

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

****Ready**.**

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

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

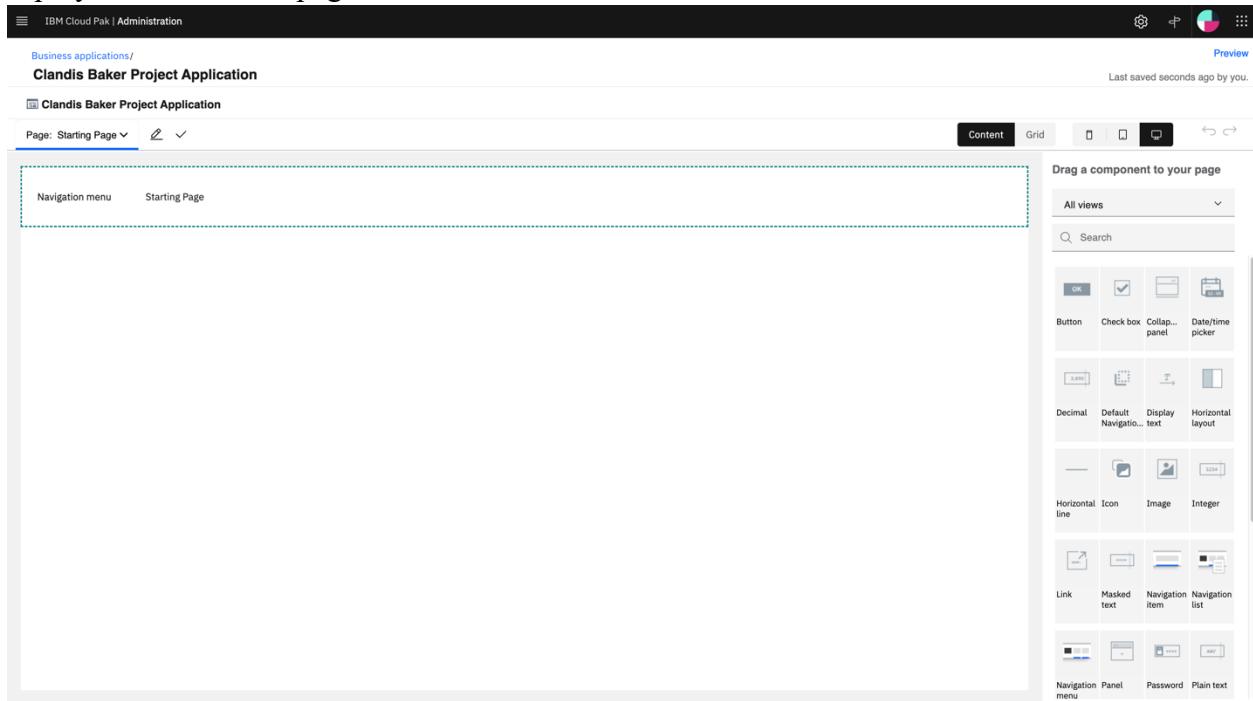
At the moment, there is a known Red Hat OpenShift bug that can intermittently block the successful deployment of CP4BA Starter pattern. To identify that your TechZone provisioned environment has hit this issue, **please check about one hour after the cluster has become ready** if your cluster is affected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

For this, please perform the following steps:

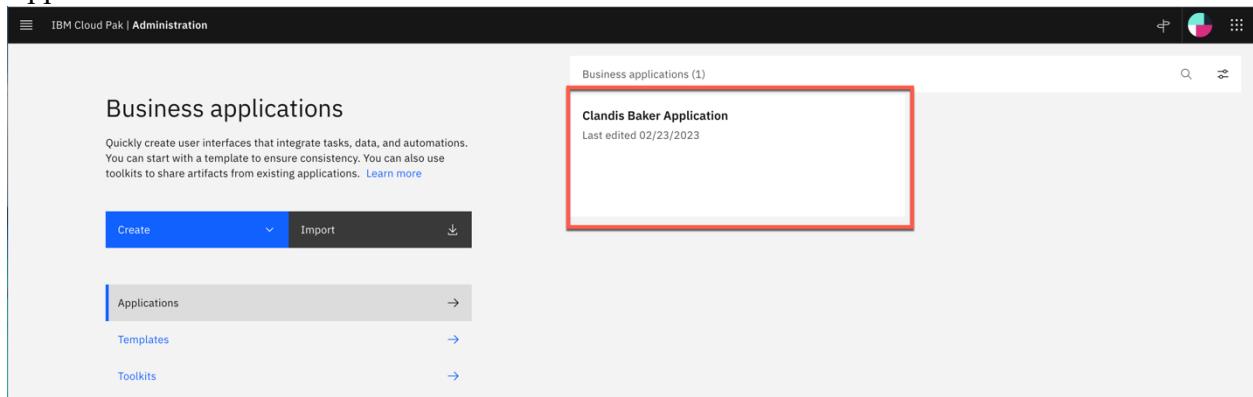
- a. Open the ****OpenShift web console**** in a browser
- b. In the left-hand side navigator go to ****Operators -> Installed Operators****
- c. Make sure the ****project scope**** is set to ****All Projects****
- d. Verify that ****all** Operators**** show in the column ****Status**** the value ****Succeeded****
- e. 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)****
- f. Once all Operators show in column ****Status 'Succeeded'****, you can proceed with the next prerequisite 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.

10.3 APPLICATION BLANK

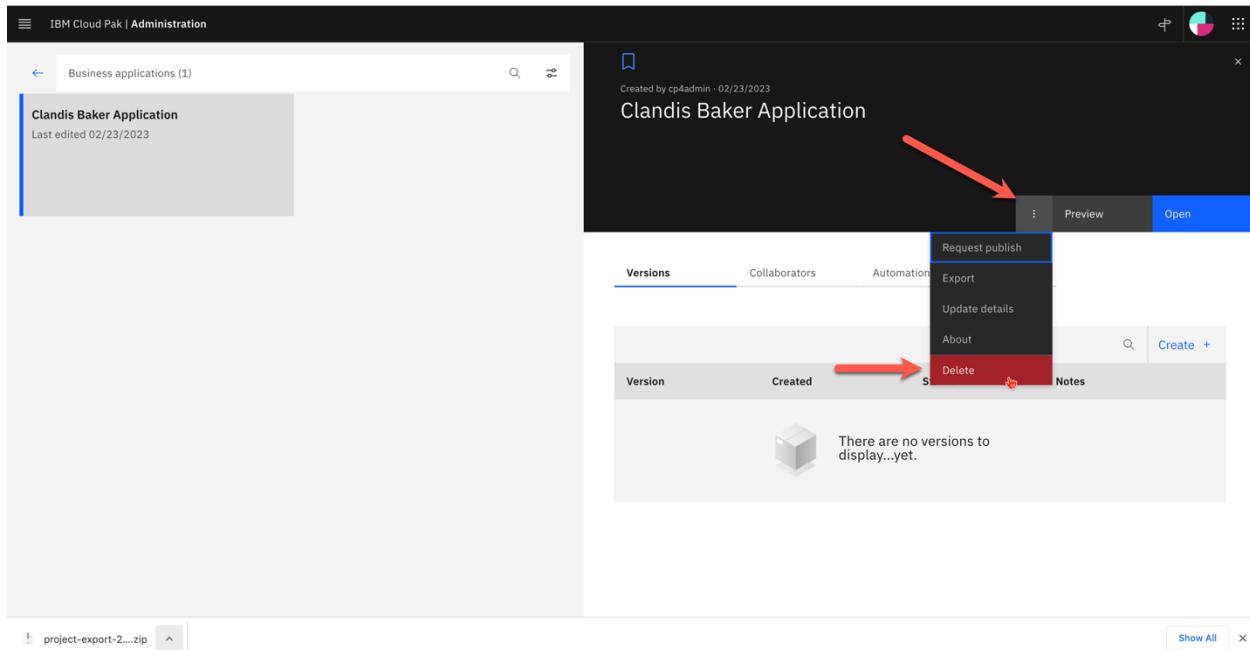
During creating of Business Application setup, sometimes on first time after project has been deployed. The Starter page is blank.



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



10.4 Connection issue with Workstation to Cloud.

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

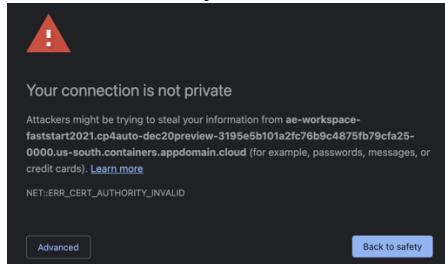
10.5 OPENING AN INCOGNITO WINDOW

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

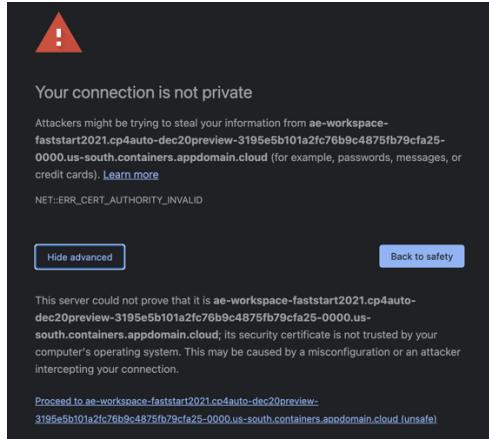
In your incognito window, go to the following URLs located in this Box:

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

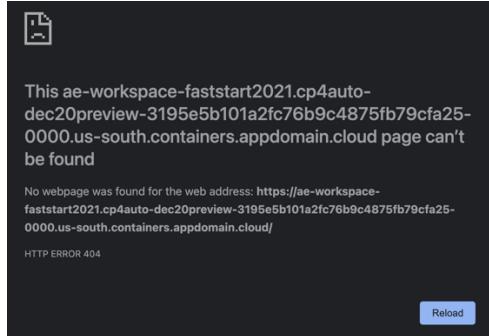
For each URL, your browser window will show a message like this:



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



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



Ignore the error and proceed to the next link.

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