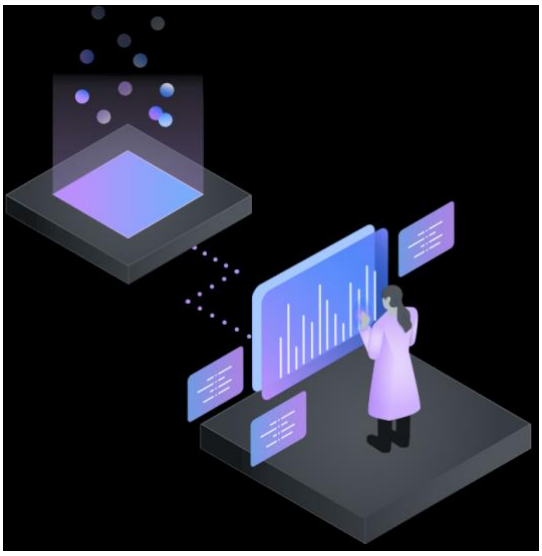
IBM Journey to Cloud and AI

Analytics Modernization Workshop

Featuring: IBM Cloud Pak for Data

Core Lab Workbook



Lab workbook and IBM Cloud Pak for Data workshop design and environment by:
Burt Vialpando, IBM Executive Analytics Architect
Kent Rubin, IBM Solution Architect

October 26, 2020

Acknowledgements

- **Duane Almeter** and **Eric Watson** for their leadership on the project
- **Daniel Kikuchi** for the CPD 3.0 cluster install environments, and other technical input and support
- **Ed Duhe, Rich Russo** and **Mitchell Odum** for providing the ESX server development platform
- **John Lucas** for product testing, workbook publishing, and project management support
- **John Van Buren** for the Organize lab designs
- **Rajesh Kartha** for Data Virtualization caching lab and DV z/OS lab
- **Benjamin Herta** for the AutoAI non-AVX work around
- **Sidney Phoon** for the last Notebook updates
- **Eric Martens** for the OpenScale lab design
- **Owais Hashmi** for OpenScale storyboarding and lab assistance
- **Rohit Gargate** for OpenScale Auto setup assistance
- **Ben Chard** for providing Analyze lab resources and assistance
- **Tom Konchan** for SPSS/IPS and Decision Optimization lab
- **Daniel Hancock** for the NPS Getting Started lab
- **David Trotter** for the z/OS work on the DV z/OS lab
- **Linda Snow** for providing assistance with complete workshop review

Table of Contents

[This page left intentionally blank]

Lab 01 GETTING STARTED

1.1 IBM Journey to Cloud and AI: Analytics Modernization Workshop

This workshop provides hands-on experience with Cloud Pak for Data that will show you how to modernize your microservices applications by enriching them with Machine Learning (ML) and Artificial Intelligence (AI).

The Journey to AI requires a strong information architecture (IA) that supports self-service capabilities and balances the needs of both the agility required by lines of business as well as the "Enterprise class" delivery required by IT. This journey can move significantly faster and with more efficiency when you use a single integrated platform like Cloud Pak for Data. It is the world's leading platform that allows you to **Collect**, **Organize** and **Analyze** data, and then **Deploy** the results to **Infuse** your applications with AI.

1.2 Cloud Pak for Data: defined

Cloud Pak for Data (CPD) is an integrated end-to-end data and analytics platform designed to help make data more accessible and trusted, as well as to provide access to many analytical tools to help your organization gain insights from your data.

CPD provides the data platform that accelerates the journey up the "AI Ladder." With it, you can quickly build, train, deploy, and manage machine learning (ML) models to create applications with Artificial intelligence (AI). CPD provides inventory and cataloging of your data sources, self-service shopping for data, and data integration and refinement capabilities. Thus, high quality and trusted data can be more easily prepared, assembled and used in one modern, integrated, collaborative and scalable platform.

Cloud Pak for Data is installed on the foundation of OpenShift for the cluster this workshop uses.

1.3 OpenShift: defined

Red Hat OpenShift is an open, hybrid cloud Kubernetes platform used to build, run, and scale container-based applications. OpenShift itself is built upon a foundation of Red Hat Enterprise Linux. OpenShift includes everything you need to manage your development lifecycle, including standardized workflows, support for multiple environments, continuous integration, and release management.

Cloud Pak for Data can be installed and managed on public cloud platforms as well, including IBM Cloud, AWS, Microsoft and more.

1.4 Lab workshop environment

We are using a CPD cloud cluster for this workshop. This software environment was built with the following key software components:

- `Red Hat OpenShift Container Platform (RHOCP) 4.3.21` as the foundational cloud-native technology platform of Kubernetes and Docker, as well as other open-source tools.
- `Cloud Pak for Data v3.0.1` as the microservice-built, integrated data and analytics platform with various add-ons installed and enabled.

1.5 Audience for this IBM workshop

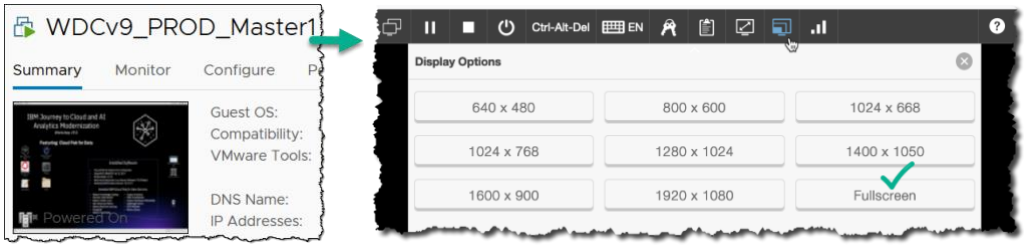
This IBM workshop is aimed at the line-of-business professionals who are tasked to gain new insights from all available data – regardless of its type and origin. The following personas who will be represented in the various labs will greatly benefit from this workshop:

Persona (Role)	Capabilities
 Administrator	Administrators set up and maintain the CPD environment itself. Note: while some of the Admin work can be done in the CPD web client, most of the Admin work on the cluster would be done in OpenShift which is outside the scope of this workshop. `The exercises in this first lab represent some typical CPD Administrator activities.`
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.
 Data Steward	Data Stewards bring integration and transformation of the data as well as providing governance, lineage and classification of the data.
 Data Quality Analyst	Data Quality Analysts perform advanced curation of the data and analyze the quality of the data.
 Business Analyst	Business Analysts deliver value by analyzing data to answer questions and communicating the results to help make better business decisions.
 Data Scientist	Data Scientists bring expertise in statistics and the process of building ML/AI models to make predictions and answer key business questions.
 Developer	Developers create and maintain the end-user applications that utilize the output from all the other personas on the CPD platform.

1.6 Getting started

__1. To launch your CPD cluster, you will select the virtual machine with the label `Master1`.

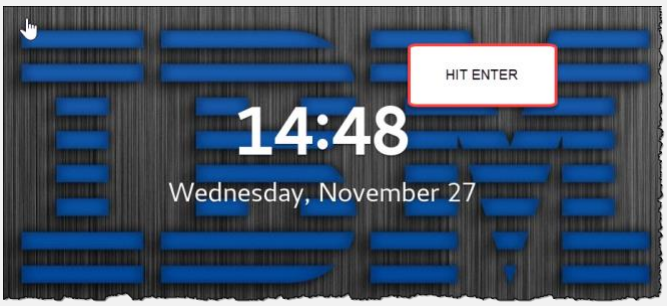
Click on this node (VM) to expand the image and then click on the `display` option in the top bar and select `Fullscreen`.



You are now in the first (headed) virtual machine image of a cluster of virtual machines that comprise your Cloud Pak for Data cluster. Everything you will be doing in the labs will be driven from this first image.

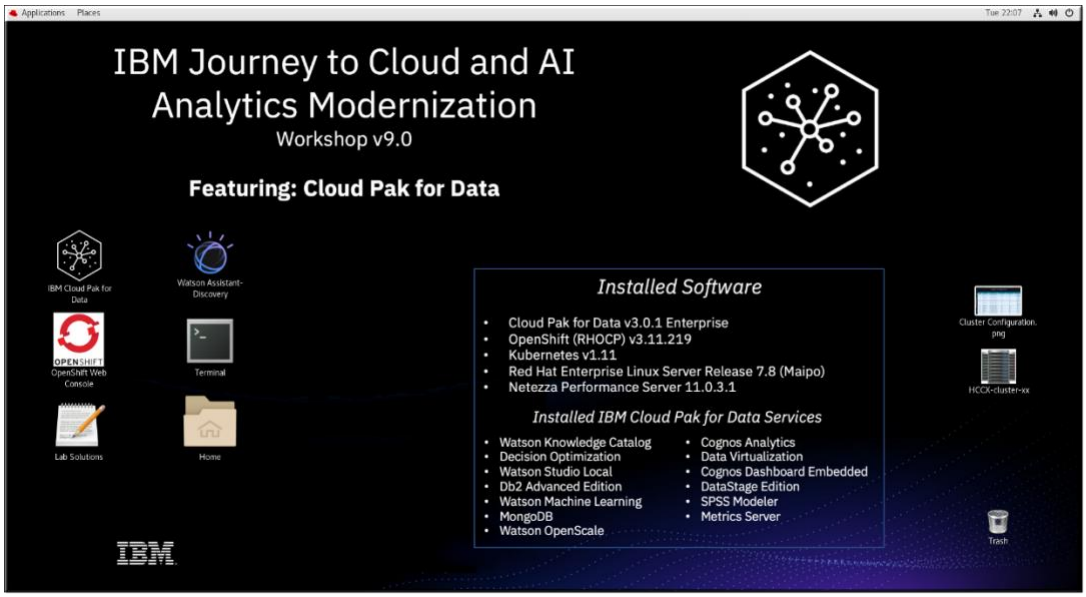
Admin

If a screensaver function has locked the screen, hit `[Enter]` to get to the desktop.



If necessary, you can log back into the Linux OS with: User `ibmdemo`, Password `passw0rd.`

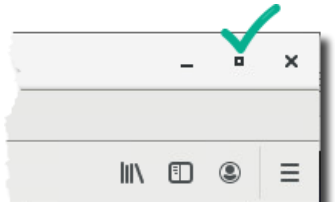
__2. In this VM, notice the lab desktop looks like this:



__3. Make sure to run in “Fullscreen” mode to make the most of your computer screen’s real estate.

__4. Double-click the icon IBM `Cloud Pak for Data`.

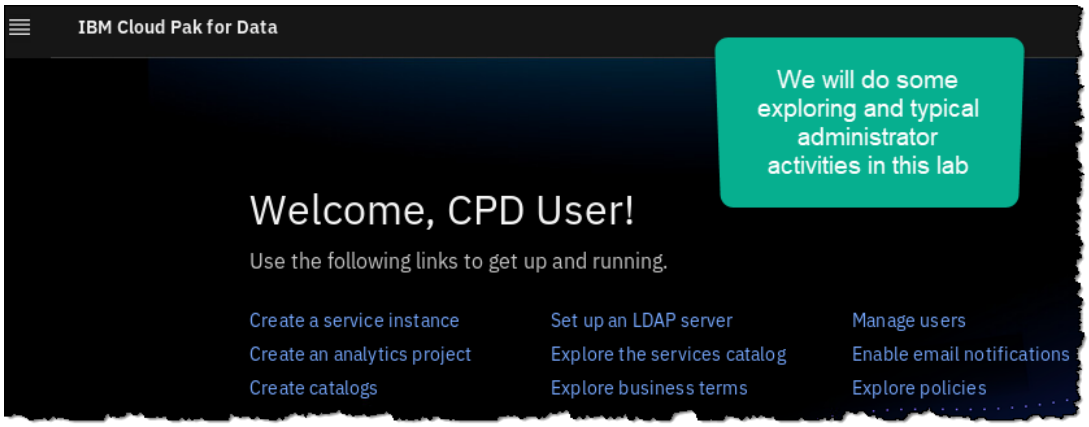
__5. After launching, maximize the browser window.



__6. The CPD web client GUI displays as shown below.

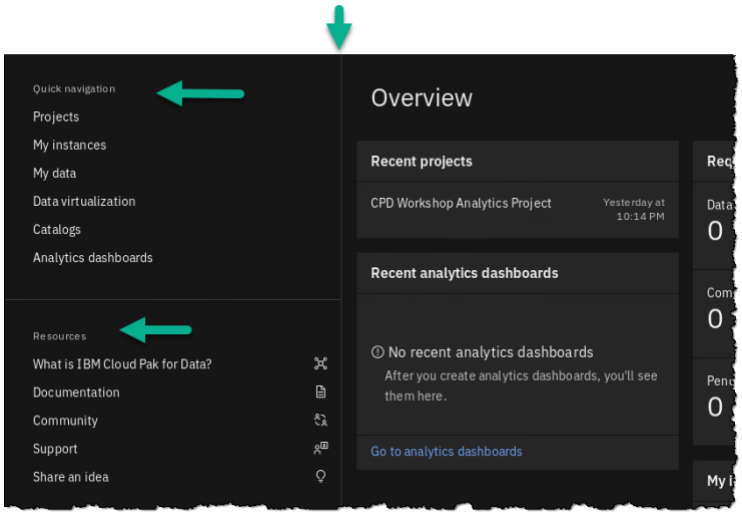
Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

__7. You should now be at the `Home Page.`



__8. Scroll down to review the quick navigation and resources links on this page.

You will be exploring many of these in more detail in this workshop, so don't follow these links right now.



__9. If you do happen leave this page by clicking on a link and you want to return to it, you can do so by clicking:

`Navigation Menu` ("hamburger" icon) ⇨ `Home`

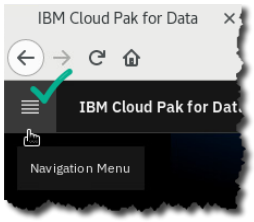
__10. You can, of course, also use the browser back and forward arrow keys to navigate through main screens in the CPD web client.

Admin	In this workshop, we will demonstrate the `Collect`, `Organize` and `Analyze` capabilities to create a machine learning model that you can `Deploy` and then `Infuse` into a microservices application.

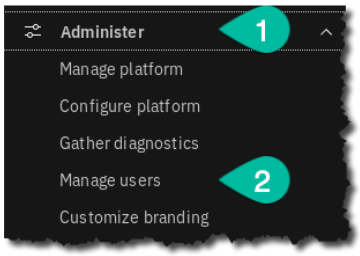
1.7 User management: Persona-based roles and teams

This section explores user authorizations for the various stages of the data analytics pipeline.

__11. Click on the `Navigation Menu` (“hamburger” icon) at the top left of the screen.

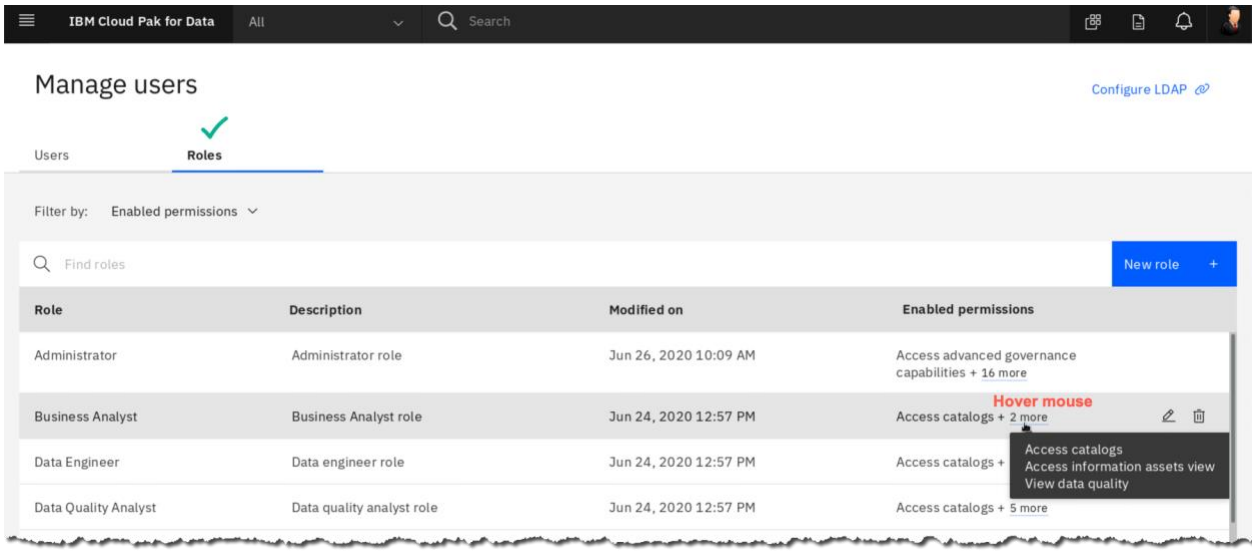


__12. Click on the `Administer` Group (to display the drop-down menu) then click `Manage users.`



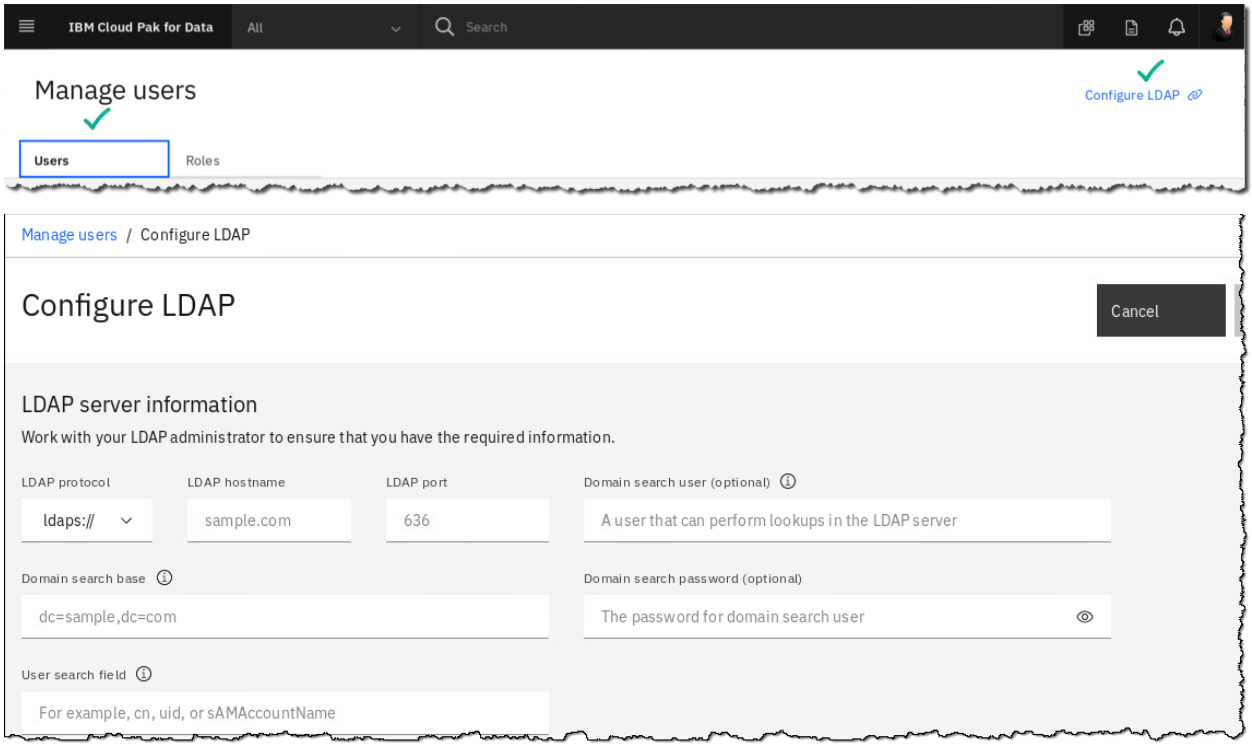
__13. Review the `Users` first, then click on the `Roles` section to review the various personas that can be represented by any given user. A user can be granted more than one role if needed.

`Hover` over the permissions of any role to see what individual permissions exist for that role.



 Admin	These existing roles can be edited (customized) or new ones created, to suit your organization’s needs. CPD is very much “persona driven” in that each user can play their particular part in your organization’s journey to AI. Each user (acting as one or more personas) can hand off and/or share their work with other users/personas, for a totally collaborative environment.

__14. Click back to the `Users` section and then click on `Configure LDAP`. You can review the fields required to do this here.

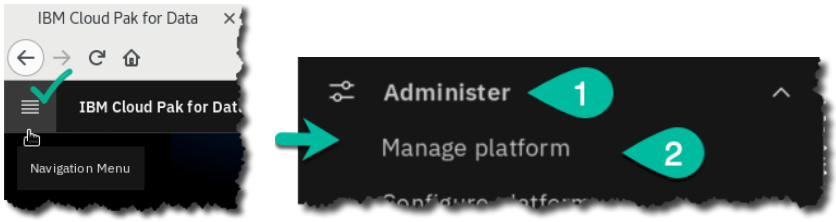


 Admin	For the sake of simplicity, you will remain logged into the CPD web client throughout this workshop as the user `cpduser`, which has been granted all persona roles. This was done so that you will not be required to log off and log on again as different users to represent the varying personas as you make your way through each lab. In your organization, however, it is likely that once you have a mature CPD environment set up, separation of duties will be defined by persona where different users will be assigned one or more personas to do their particular tasks.

1.8 Managing platform options

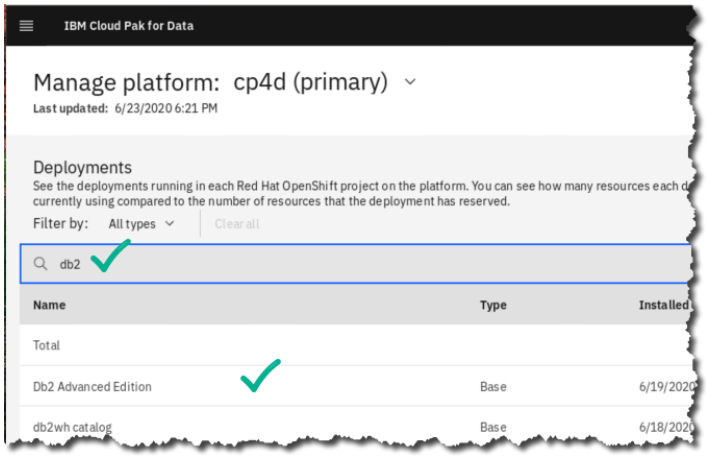
You can view the underlying OpenShift services and pods by doing the following:

__15. Click `Navigation Menu` ⇨ `Administer` ⇨ `Manage platform.`



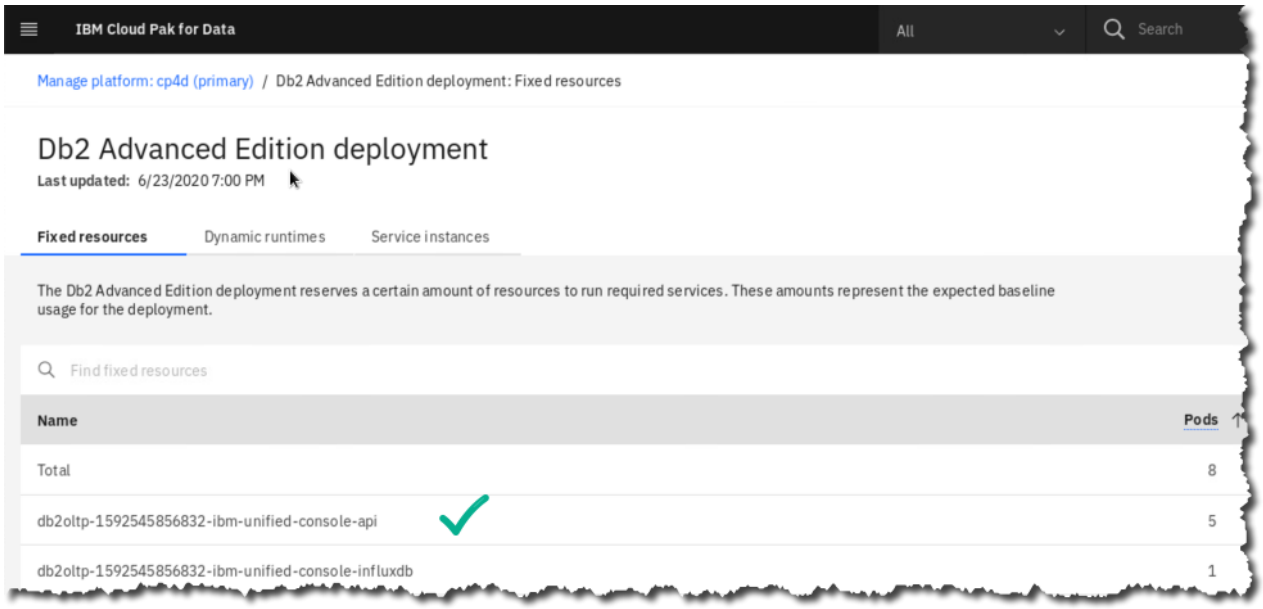
__16. In the Search area, type `db2.`

__17. Click the deployment link `Db2 Advanced Edition.`



__18. Under the tab `Fixed resources`, notice the Deployment CPU and Memory usage.

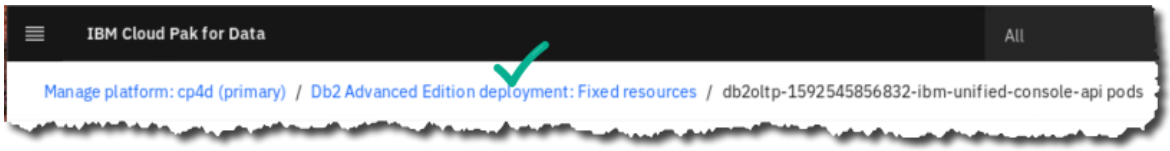
Click on any Deployment to review the pods for it.



__19. The next screen shows the underlying OpenShift/Kubernetes pods for this deployment. (Your pod names will differ.)

Pods	SCC	Service account	Created on	Status	vCPU
Total					0.19 of 1.00
db2oltp-1592545856832-ibm-unified-console-api-599c57d76...	db2u-scc	db2u	6/19/2020 1:51 AM		0.19 of 1.00
db2oltp-1592545856832-ibm-unified-console-collector-1592...	db2u-scc	db2u	6/23/2020 7:05 PM		0.00 of 0.00
db2oltp-1592545856832-ibm-unified-console-collector-1592...	db2u-scc	db2u	6/23/2020 7:03 PM		0.00 of 0.00

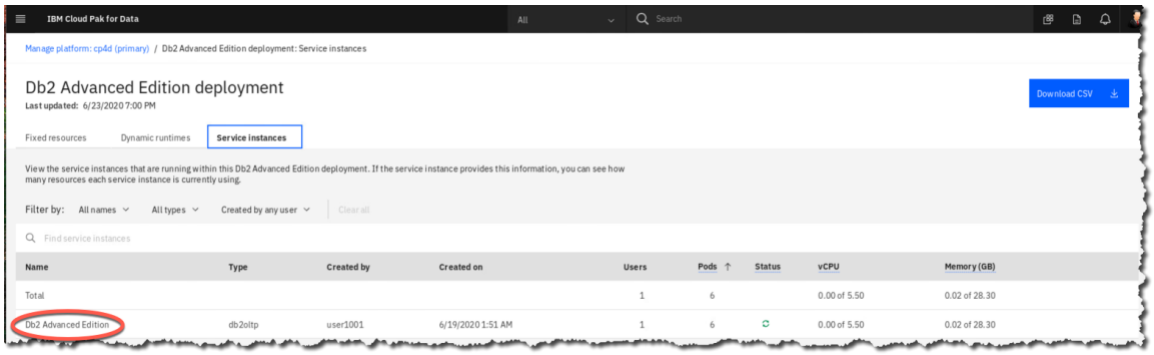
__20. Click on `Fixed resources` in the bread crumb line (or use the back arrow).



__21. Click on tab `Service Instances.`

This shows how many instances of the Db2 Advanced Edition have been created using the Db2 Advanced Edition deployment. In our case, it is only one.

We will review this instance in a different way later in this lab.

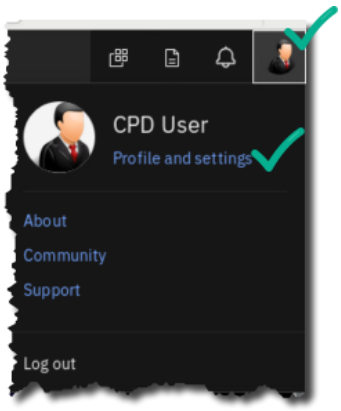


1.9 Reviewing the profile settings

__22. Click the top right circle of your web client screen that has your user icon on it. This provides a drop down.

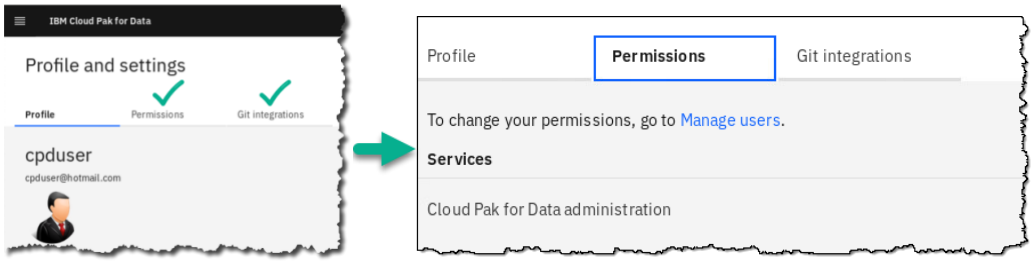
Choose `Profile and settings.`

(Note: this is also the location where you can `Log out` of the web client)



__23. Review the things you can change in your `Profile`, then review `Permissions.`

Note: in Permissions, your user has all permissions to allow you to do anything required in the workshop lab exercises. In the real world, your permissions would be more limited and controlled by an administrator.



__24. Now review `Git Integrations` ⇨ `Generate API key` and `New token.`

These allow you to configure Git with CPD, which allows you to integrate CPD projects with your current CI/CD (Continuous Integration and Continuous Delivery) pipeline to automate delivery of the artifacts you create in the CPD platform. You can use the capabilities from the underlying OpenShift platform to build cloud native microservice applications which are tied to the ML / AI model development with a delivery pipeline.

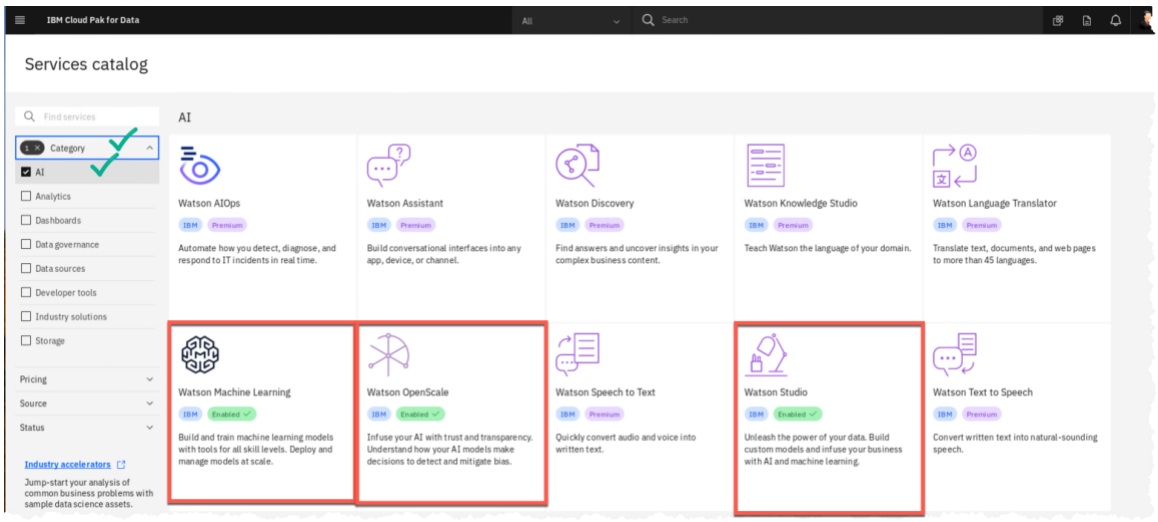
1.10 Reviewing CPD services

__25. Click the `Services` icon (four little squares over one bigger square) on the top right corner of your screen.

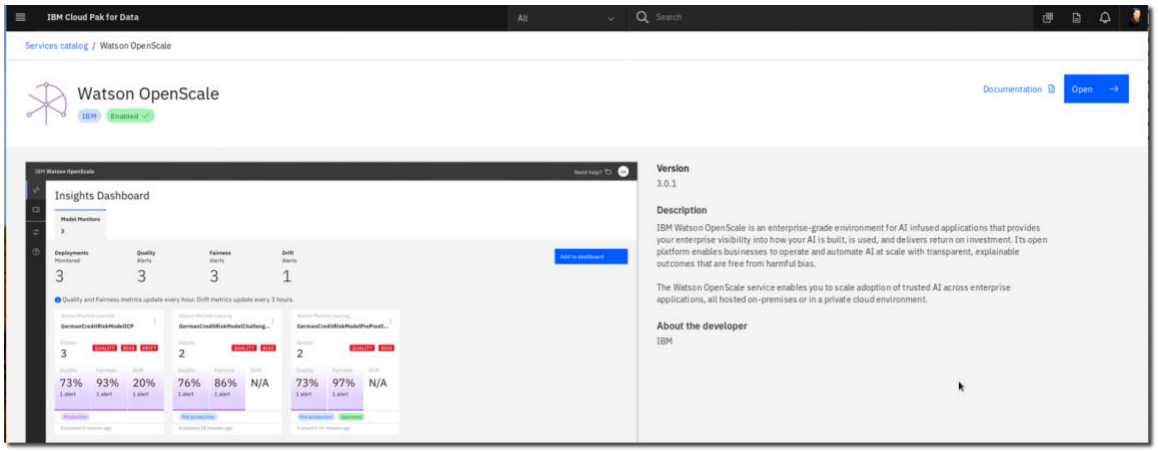
__26. This will bring up all available services for Cloud Pak for Data.

Click through the various categories to see what services you can install on CPD.

For example, click on category `AI` and notice what is available there. Those with “Enabled” are accessible by this CPD cluster right now.

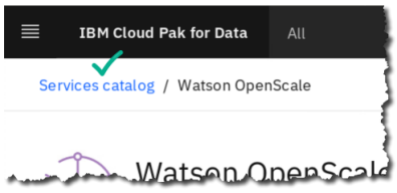


__27. Now, click on the `Watson Machine Learning` service tile to get more details on what this service can do.



__28. Click the browser back arrow to return to the `Services catalog` screen.

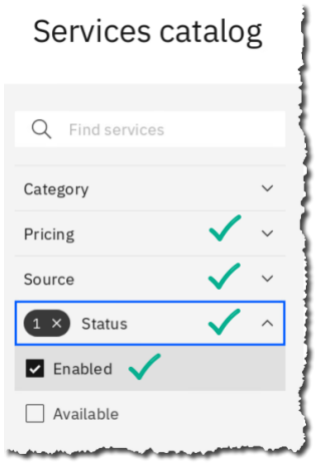
(Or you can click on the `Services catalog` link in the breadcrumb trail itself)



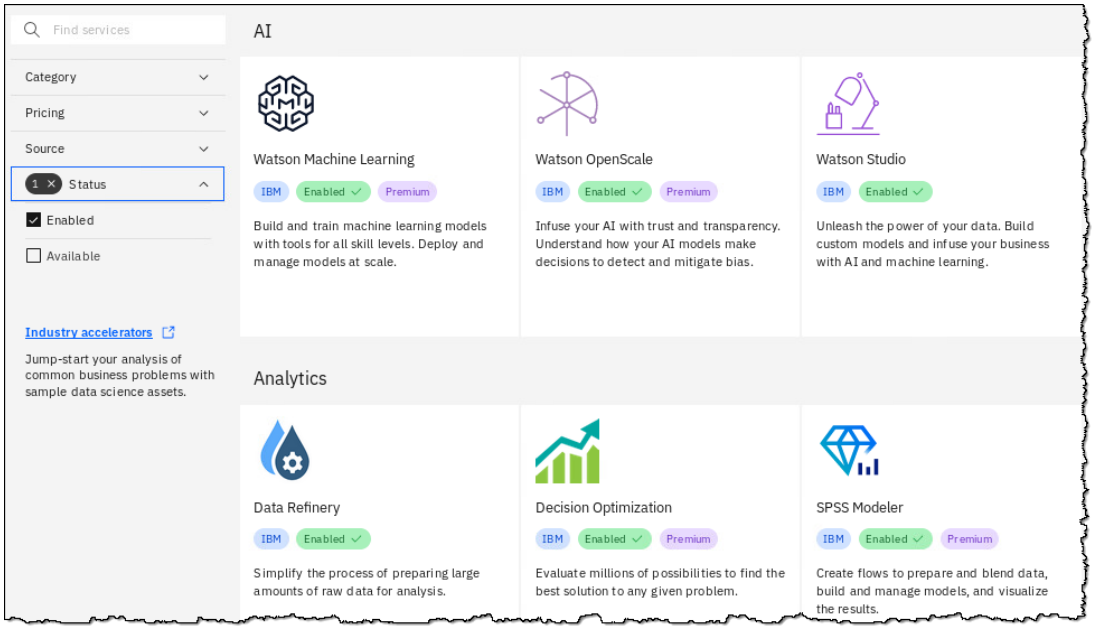
__29. Another convenient way to review these Services is to filter them in the Services Catalog options.

Click through `Pricing`, `Source` and `Status`.

In `Status`, check `Enabled`



__30. This will show all the enabled services in your cluster (as long as you did not check a filter for Categories, Pricing or Source.)

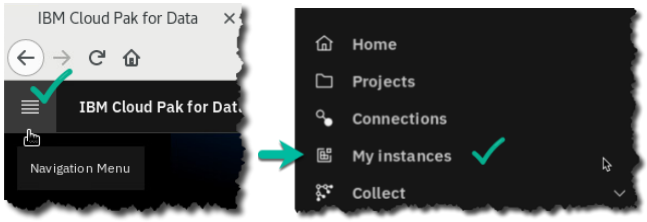


__31. Explore through this Services page to find a few capabilities that you might find useful for your organization.

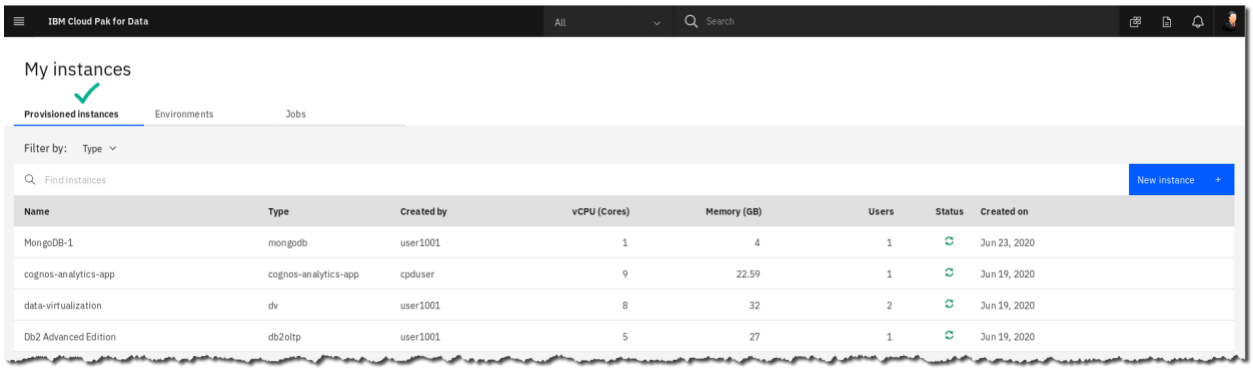
Admin	IBM continues to provide more services with each release of Cloud Pak for Data. Some are part of the base offering, others are purchasable as IBM "cartridges," and still other are purchasable through a 3rd party vendor.

1.11 Reviewing instances

__32. Click `Navigation Menu` ⇨ `My instances.`



__33. Click tab `Provisioned Instances` and then click the twistie to sort the instances that were provisioned for this workshop.



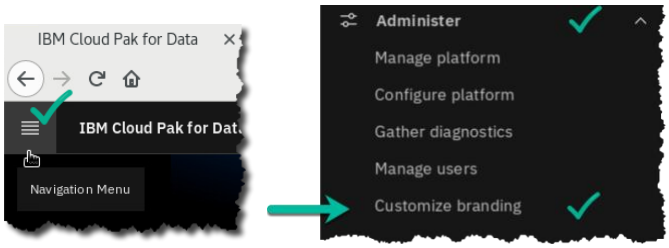
Admin	The term "instance" in this context means a copy of a persistent data store within the CPD platform. These instances are stateful Kubernetes services like Db2, MongoDB, Streams, Data Virtualization and even Cognos Analytics.

__34. Check the tabs for `Environments and Jobs.` Note: the cluster may not have any of these running at this time, so these pages could be empty.

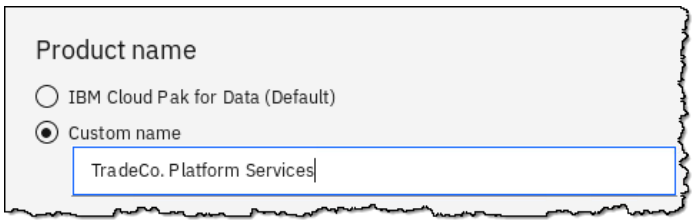
Admin	The term "environment" in this context means a copy of an analytics runtime that is running (taking up resources) on the cluster. These can be Jupyter/Python, Zeppelin/Anaconda, R Studio, Decision Optimization, etc. The term "job" in this context means a task scheduled within the platform. These can be analytics related (like a scheduled batch scoring job) or they can be a scheduled ETL or Streams job, etc.

1.12 Customizing branding

__35. Click `Navigation Menu` ⇨ `Administer` ⇨ `Customize branding.`



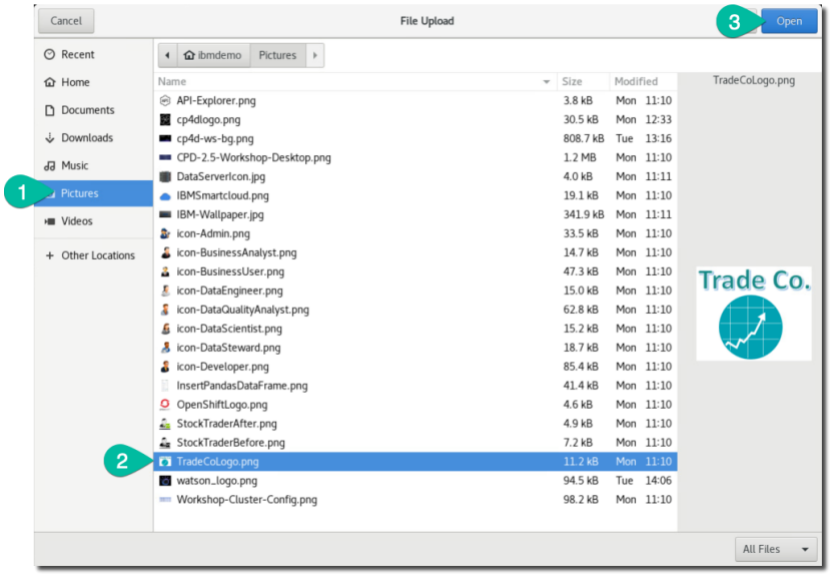
__36. In section `Product name`, click `Custom name`, then fill in `TradeCo. Platform Services.`



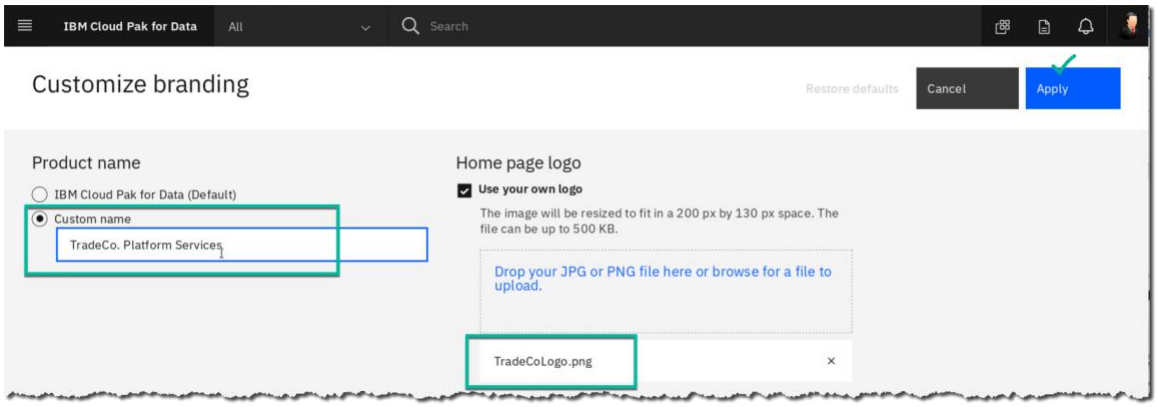
__37. In section `Home page logo`, check `Use your own logo`.

Click on the box: `Drop your JPG or PNG file here or browse for a file to upload.`

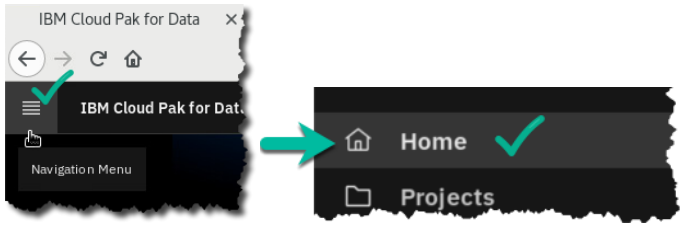
__38. For Browser Only: Download TradeCoLogo.png from https://ibm.biz/BdqhHa.
For Unified Desktop: Under directory `Pictures`, select file `TradeCoLogo.png` ⇨ `Open.`



__39. Click button `Apply.`



__40. Click `Navigation Menu` ⇨ `Home.`



__41. Notice the Navigation bar is customized, and the Home Page can have a company logo on it.
Note: You may have to refresh your browser if it does not immediately show.

__42. Let's reset this to the default setting so the rest of the lab workbooks will be consistent with your environment:

Click `Navigation Menu` ⇨ `Administer` ⇨ `Customize branding,` then `Restore defaults` ⇨ `Restore defaults.`

1.13 Lab conclusion

Cloud Pak for Data is useful in the following business macro use-case scenarios:

1. **Manage Your Data Anywhere**: Use data virtualization, streaming, cataloging, governance and more to prepare your data for analysis.
2. **Operationalize Data Science & AI**: Build, deploy, manage & govern models and data at scale to improve business outcomes like controlling customer churn, cross selling and up selling, predictive maintenance and more.
3. **Shift to Next-Gen workloads**: Shift to Cloud Native to be able to provision and scale in minutes, build once and deploy anywhere with multi-cloud support, and use built-in automation and collaboration to increase productivity.
4. **Smarter Governance:** Enable self-service analytics with auto-discovery of meta data, implementing governance rules and policies, enforcement of privacy to mitigate risk and to ensure compliance for regulatory requirements like GDPR.

Make Cloud Pak for Data your platform for data and analytics. Why? Because IBM understands data and provides an integrated, end-to-end data platform that enables enterprises to:

- Collect relevant data and make it simple and accessible
- Use federation, virtualization and/or transformation to combine and refine data sets
- Organize data so it can be trusted
- Analyze insights on demand
- Infuse machine learning into your applications

All of the above and much more will be demonstrated in the following workshop labs.

**** End of Lab 01 – Getting Started**

Lab by Burt Vialpando and Kent Rubin, IBM

Lab 02 BUSINESS USE CASE: CUSTOMER CHURN

2.1 Lab overview

Trade Co. is experiencing a decline in review, which appears to be due to losing customers. The current process of predicting customer churn seems to be ineffective as well. The company's executives have asked their senior Business Analyst to help them understand why and to help find a solution.

2.2 Persona represented in this lab

When embarking on machine learning projects, many organizations engage Business Analysts to help gain insight into their data. This persona can use tools like Analytics Dashboards to build visualizations to help the organization more clearly understand their business challenges.

The Business Analyst persona is the likely role to perform the exercises in this lab.

Persona (Role)	Capabilities
 Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.

2.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Double-click the desktop icon: `Cloud Pak for Data Web Client.`

__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

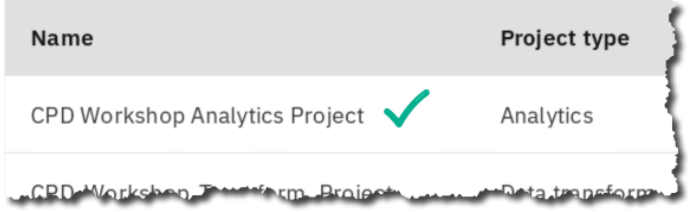
2.4 Reviewing the dashboard: Monthly Metrics

The Business Analyst has used the Cloud Pak for Data Analytics Dashboarding service to analyze Trade Co. issues. He has come up with a Monthly Metrics dashboard that demonstrates the company's concerns.

This first dashboard was built with the recent trading information delivered to him for the year. It was placed into a CSV file in the project the Business Analyst shares with his team.

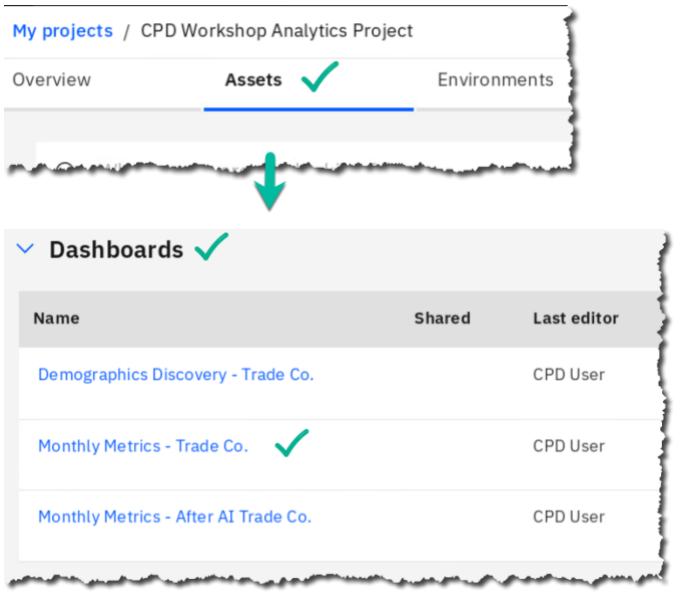
__4. In the CPD web client, click the `Navigation Menu` ("hamburger" icon) ⇨ `Projects.`

__5. Select the project: `CPD Workshop Analytics Project`.



__6. Under tab `Assets`, scroll down until you find `Dashboards.`

Click `Monthly Metrics – Trade Co.`



Business Analyst

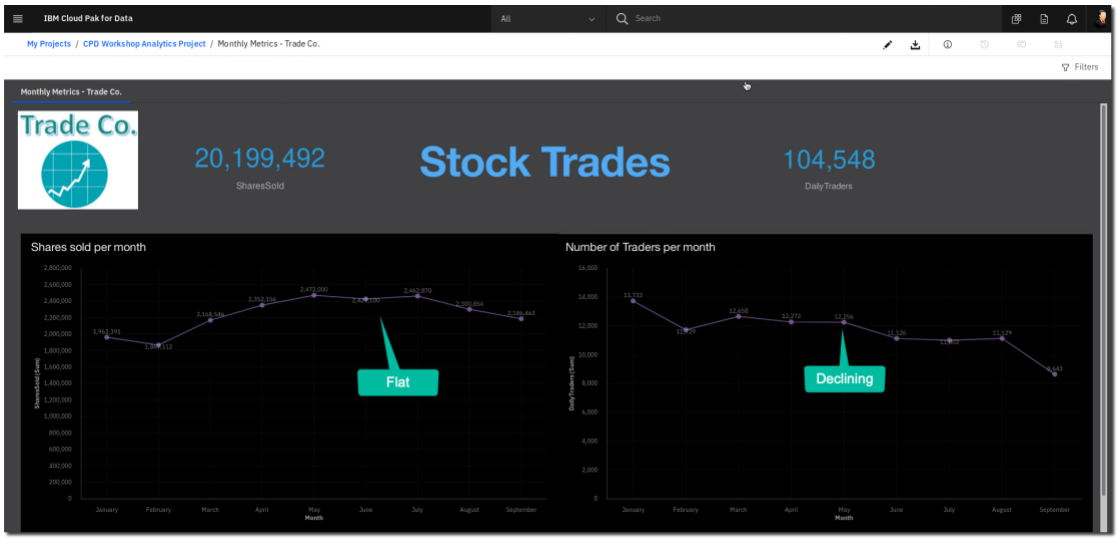
Use `[Ctrl] –` and `[Ctrl] +` keys to adjust the zoom on the dashboard to best suit your screen.

You can also use `[Ctrl] [Mouse-Scroll-Button]`.

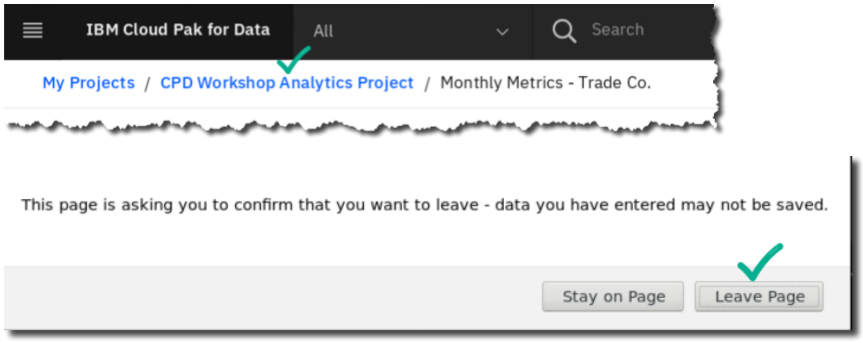
__7. Notice that from January to September, `Shares sold per month` remains flat while `Number of Traders per month` are declining.

This verifies that customer churn is in fact occurring, even though the remaining traders are trading more each month, keeping `shares sold` nearly even.

If Trade Co. could somehow find a way to retain their customers, shares sold per month would go up, thus driving revenue up.



__8. Leave this dashboard by clicking on the breadcrumb trail to navigate to the project again.



Business Analyst	If you would like to see how this dashboard was built, you can build it yourself by doing the following lab: `(Deeper Dive) – Cognos Dashboard Embedded`

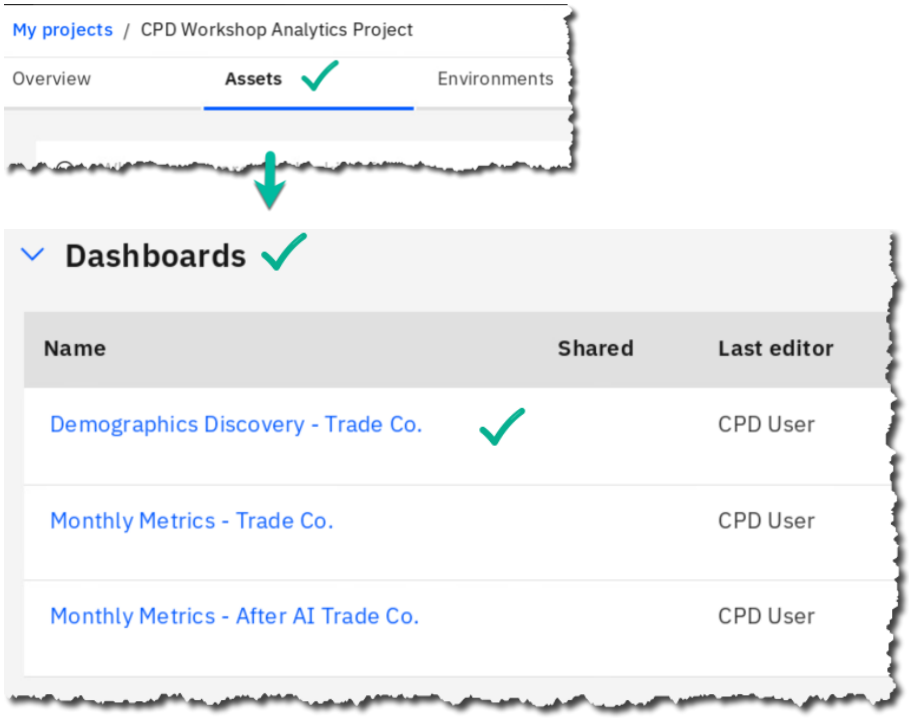
2.5 Reviewing the dashboard: Demographics Discovery

The Business Analyst next asked for the company's customer demographics data, joined together with the customer churn data, and joined again with the recent activity data.

This data (representing three different data sets together) was placed in the team project and the Business Analyst created the following dashboard to better understand the situation.

__9. Under tab `Assets` find `Dashboards.`

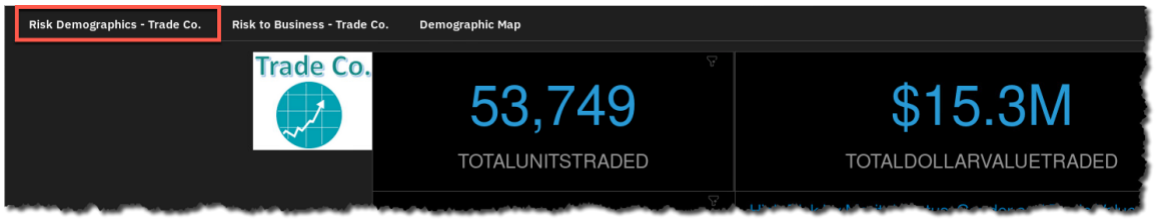
Click `Demographics Discovery – Trade Co.`



__10. Notice there are two tabs in this dashboard; you are currently positioned at tab:

`Risk Demographics – Trade Co.`

The top visualizations give `Total Units Traded` and `Dollar Value Traded`. You will see later that these are interactive with the rest of the dashboard.

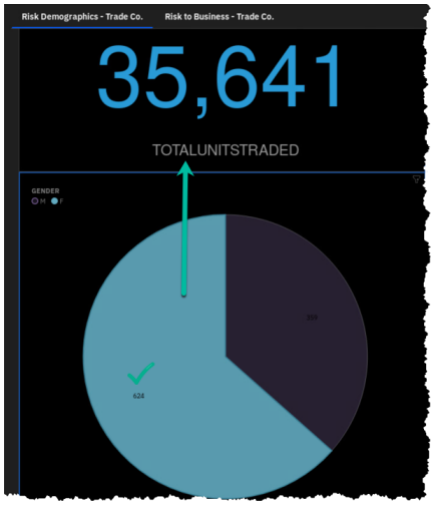


__11. Review the top left filter (funnel) icon and notice that a filter has been set to only see information for those considered `High` in `Churn Risk`.

__12. Click on the aqua blue portion of the pie chart (which represents `Female` customers).

Notice the top visualizations change in value. For example, the new number displayed in the `Total Units Traded` visualization would now indicate `High Risk Females`.

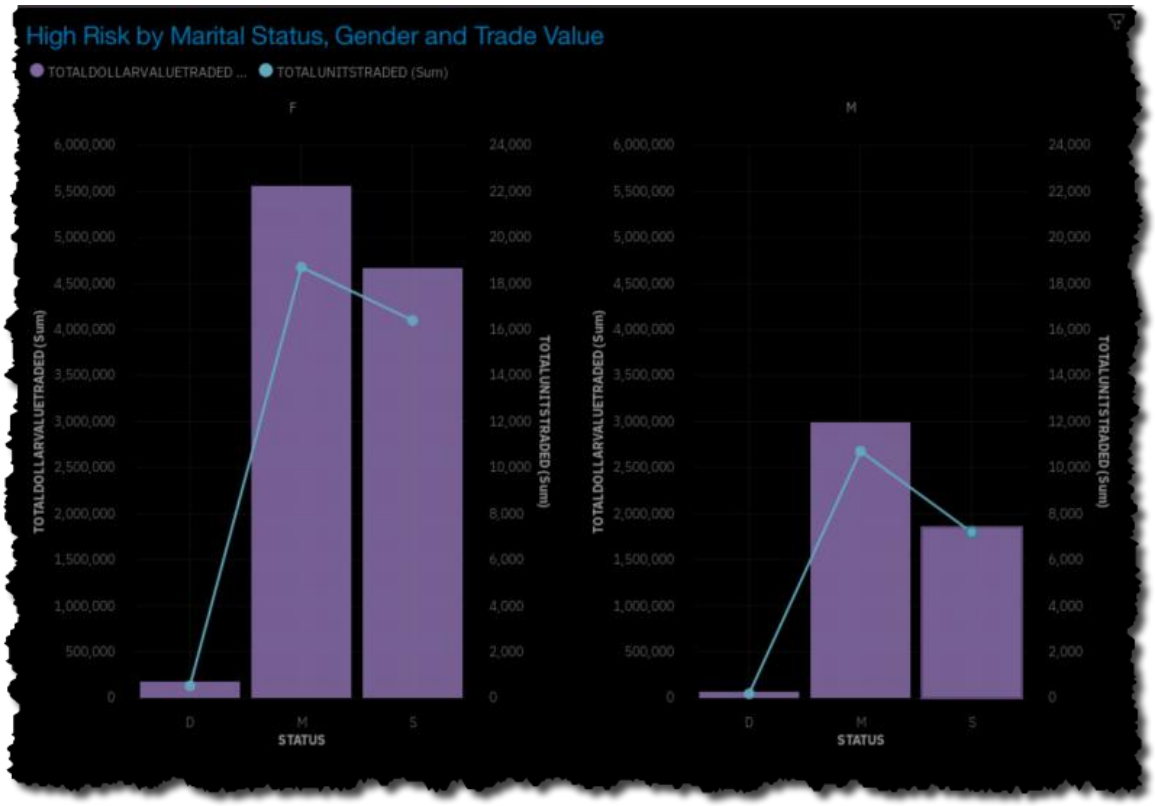
Click again on the aqua portion of the pie chart to deselect the values for `Female`.



__13. The next visualization is also filtered by High Risk, and then sub-divides the data into two charts between `Females` (on the left) and `Males` (on the right).

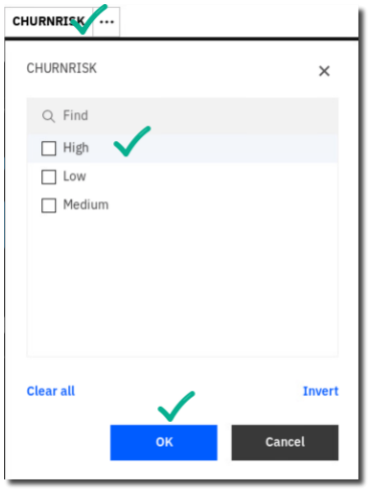
Further, each chart shows bars (`Dollars`) and lines (`Units`) for the Marital Status of Divorced(`D`), Married(`M`) and Single(`S`) customers.

This is a complex visualization, but from it you can see that Married Females are the largest group.



__14. In the top filters section, click on the filter for `CHURNRISK.`

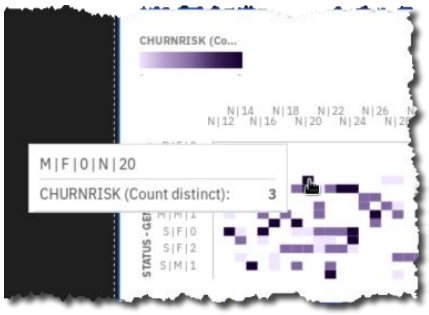
`Deselect` the filter for `High` and click `[OK].`



__15. With no filters in place for this tab on the dashboard, scroll down to the bottom visualization.

__16. The bottom visualization on this dashboard is a Heat Map of many different demographics data points along with ChurnRisk. On this map, the darker the square, the more the risk.

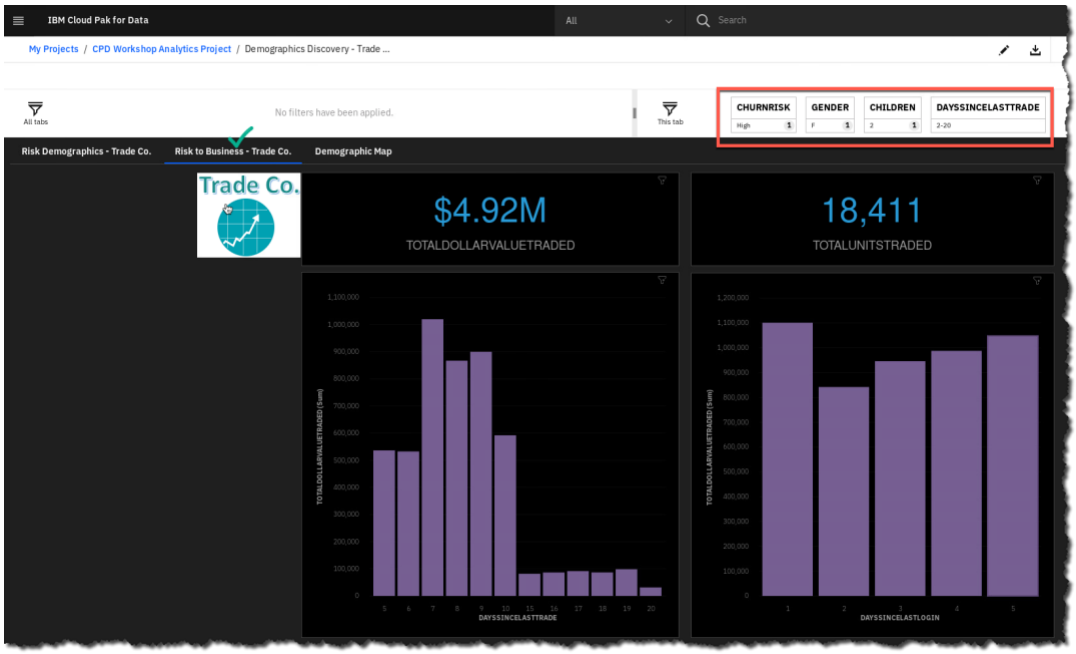
__17. Hover over darker purple square on the top left of the Heat Map. Notice it represents Married Females with zero children, who are not homeowners and are 20 years of age.



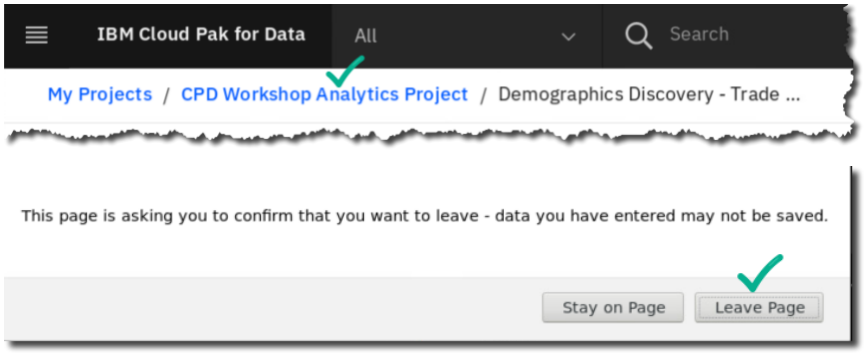
__18. Click on the second tab in this dashboard called `Risk to Business – Trade Co`.

This analyzes Dollars and Units traded on the top, filtered by the demographics you found in the first tab visualization that appear to be the higher risk. (Note: these filters can be changed at any time in Edit mode to give flexibility to the user.)

This tab also has bottom visualizations that demonstrate that even though this demographic appears to be high risk, they are still fairly active.

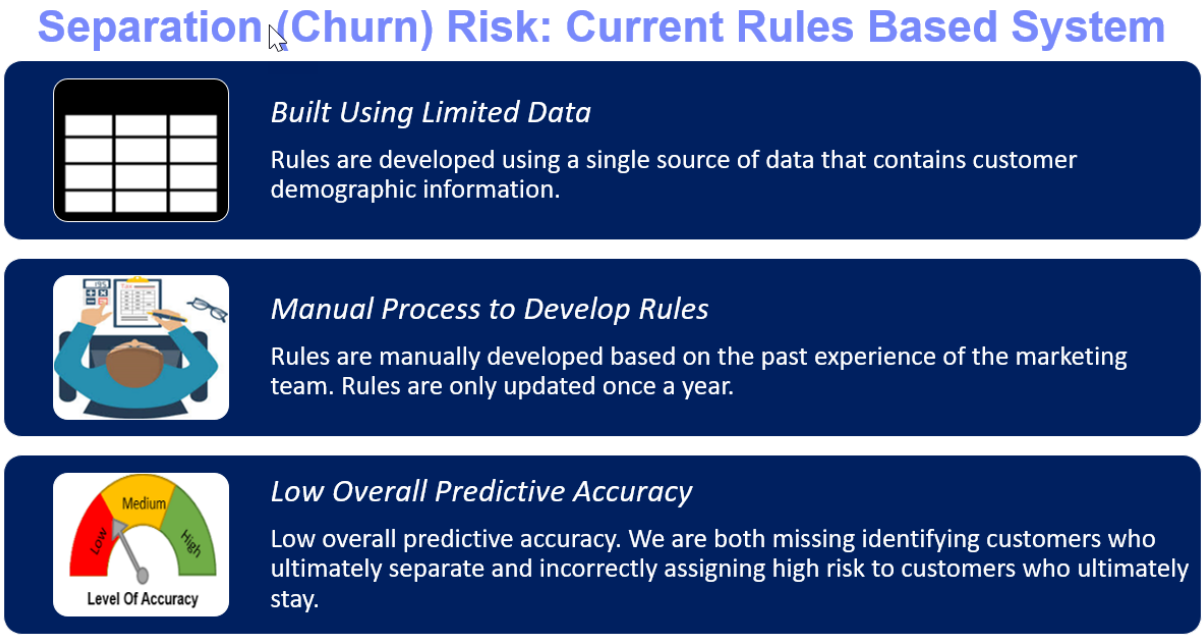


__19. Leave this dashboard by clicking on the breadcrumb trail to navigate to the project again.



2.6 Devising a plan

Armed with his findings, the Business Analyst brings this to the Data Scientists on the team. Together they come up with a plan to fix their current Rules Based system of predicting churn.



The intent is to use a better approach: Leverage Cloud Pak for Data to build a data driven Machine Learning model to infuse into their Stock Trader application.

Separation (Churn) Risk: New Data Driven Approach

Incorporate Multiple Data Sources

Use vetted centralized transactional data along with customer demographics to understand separation behavior. Also, include the outcomes of the rules based system for each customer where an accurate prediction was rendered.

Data Driven Process to Develop Machine Learning Models

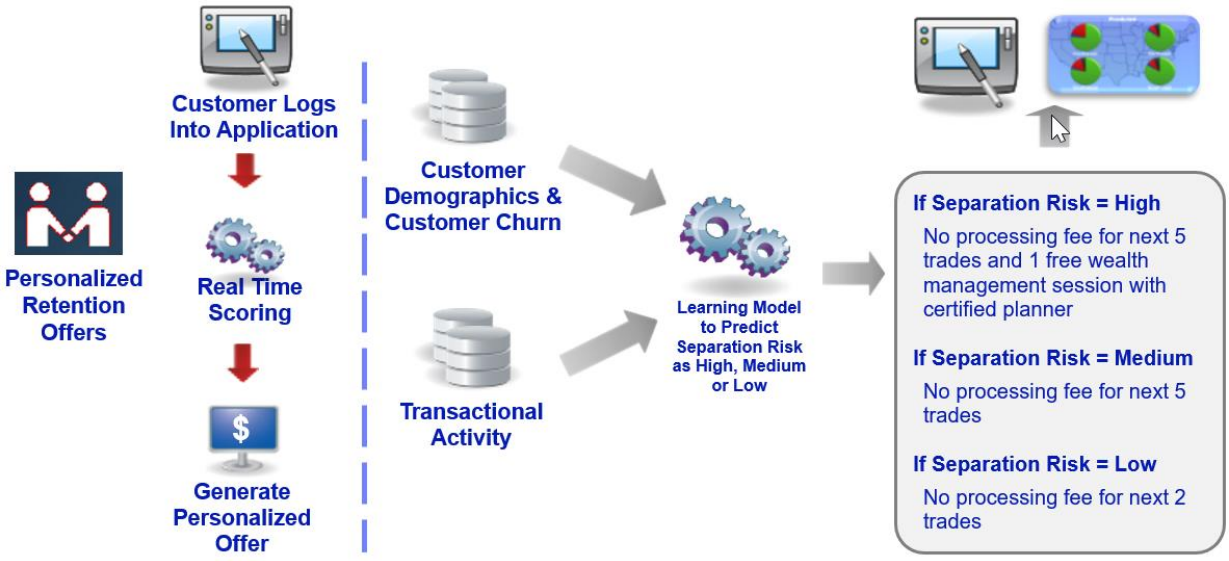
Develop predictive models for separation risk that automatically discover and incorporate all the patterns in the data including interactions and contingencies.

High Accuracy from Adaptive Machine Learning

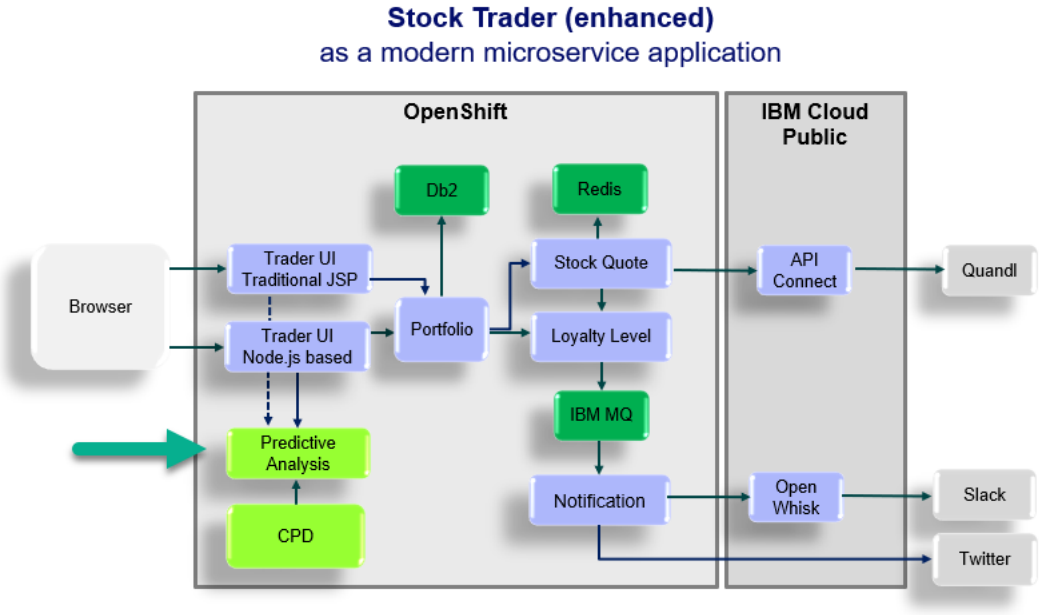
Models will classify separation risk with a higher overall accuracy and will adapt to changing patterns in risk to maintain that accuracy. Machine Learning models will incorporate all the understanding from the rules-based system and build on that to develop highly complex set of predictive conditions.

With this better model that more accurately predicts churn, the developers will infuse it into their Stock Trader application that will act upon this information in real time. Once a customer (trader) logs into the application, the new ML model will predict the risk of churn for this customer and will then make one of three offers designed to retain them.

Deployment: Stock Trader App with Integrated AI



In this case, infusing the new ML model into the Stock Trader application will be as simple as writing one extra microservice that is invoked upon sign in. It will be called the “Predictive Analysis” service (indicated by the green arrow below) and will make the offer after sign in.

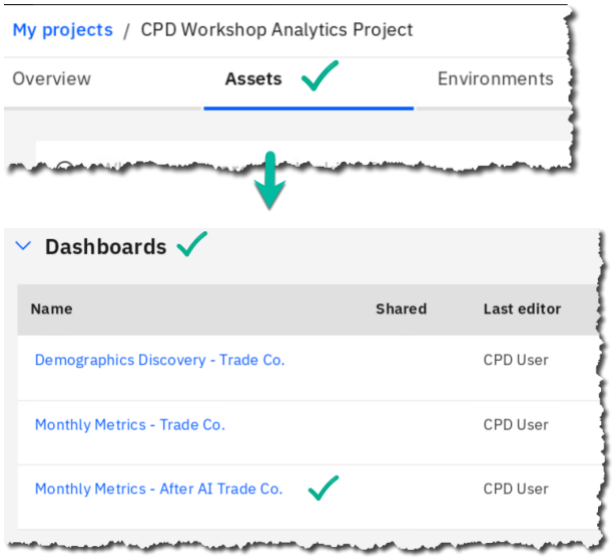


2.7 Reviewing the dashboard: Monthly Metrics After AI

The model was created and infused into the Stock Trader application, which ran for three months. New activity data was captured for this period. The Business Analyst created another dashboard similar to the first that displays the results.

__20. Still in the same project, under `Assets`, scroll down until you find `Dashboards.`

Click `Monthly Metrics - After AI Trade Co.`



__21. This dashboard is similar to the first one in the two visualizations:

`Shares sold per month` and `Number of Traders per month`

Notice however, new data is present for October, November, and December.

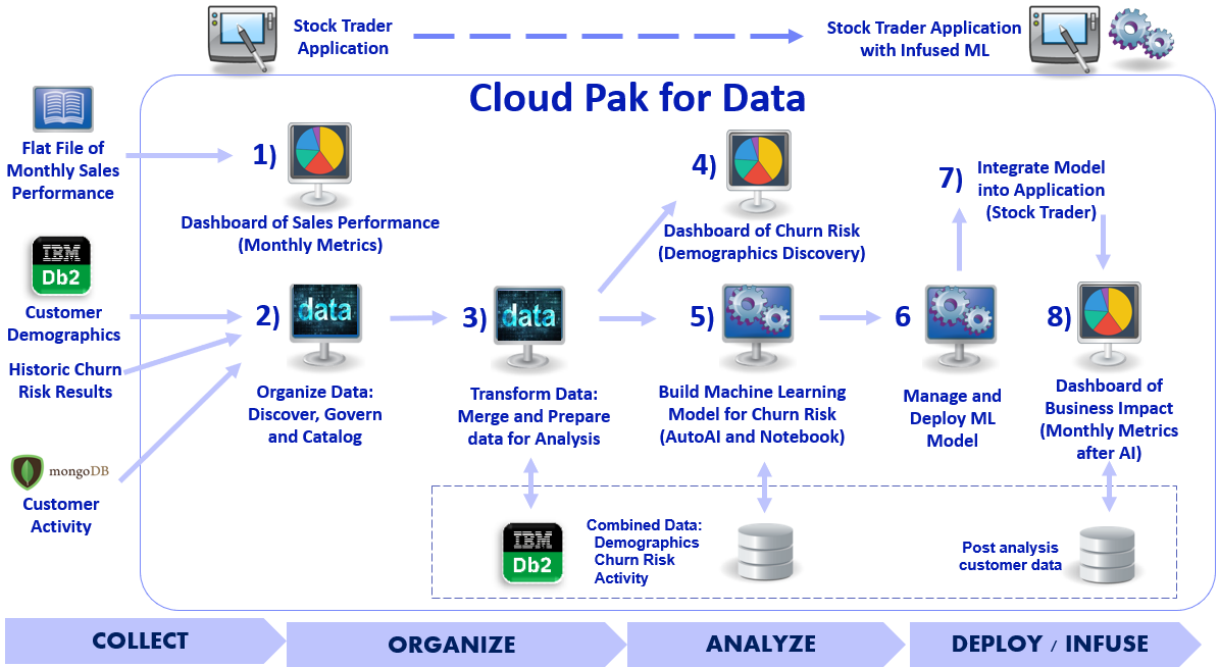
The data is very encouraging: both data points have gone up!

Trade Co. is back in business with their successful customer retention offer driven by a smarter application infused with a machine learning model that was built and maintained on Cloud Pak for Data.

2.8 Lab conclusion

This workshop will walk you through how Trade Co. was able to pull off their business success using the Cloud Pak for Data platform.

Each subsequent lab will walk you through one of the steps (Collect, Organize, Analyze, Deploy, Infuse) taken by Trade Co. in their Journey to Cloud and AI using this amazing, industry leading Analytics Modernization platform called Cloud Pak for Data. You will also take on the roles of the various personas involved along the way. The steps in the journey are depicted in the illustration below.



**** End of Lab 02 – Business Use Case: Customer Churn**

Lab by Burt Vialpando and Kent Rubin, IBM

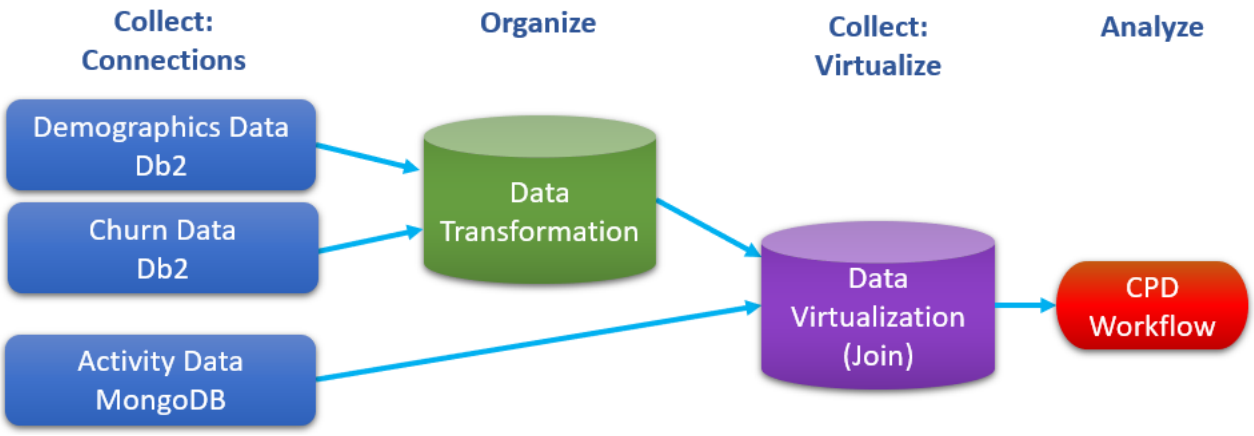
Lab 03 COLLECT: CONNECTIONS

3.1 Lab overview

The Collect capability of CPD means accessing your organization's data regardless of where it resides, whether that be an in-cluster data source (e.g. the Db2 Advanced Edition) or a native connection to a remote data source. You can even use the Db2 Event Store or Streams services to provide streaming access that is best suited for Internet of Things (IoT) processing.

Additionally, Data Virtualization and Data Transformation are available to streamline the access, performance, and formatting of the data for use in later steps of the CPD analytics workflow.

In this lab you will explore Connections for the Collect process. In later labs you will explore Data Transformation and Data Virtualization.



3.2 Persona represented in this lab

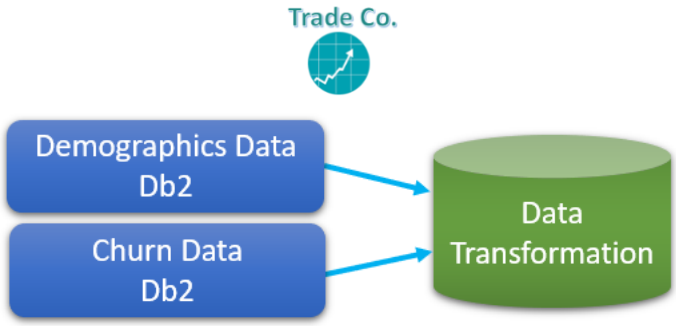
The Data Engineer persona is the likely role to perform the various Collect tasks shown in this lab.

Persona (Role)	Capabilities
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

3.3 Db2 data overview – Transforming for analytics

The Db2 data in our Trade Co. scenario (Demographics and Churn) have two key factors that make an appropriate scenario for `Data Transformation`:

- The data only changes once a month (it is relatively static). Thus, copying it and/or changing it for downstream processing in our CPD analytics workflow is OK because we are not required to have the latest data to get the results we need.
- The data must be changed before it can be processed in our CPD analytics workflow.



Although you will be doing the actual `Data Transformation` in the next lab, you will prepare for it in this one by reviewing some steps in this `Connections` lab.

3.3.1 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

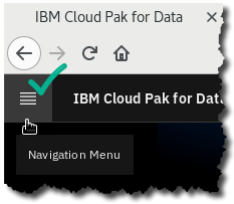
__2. Double-click the desktop icon: `Cloud Pak for Data Web Client`.

__3. The CPD web client GUI displays as shown.

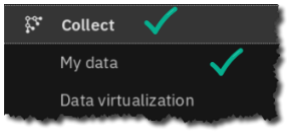
Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

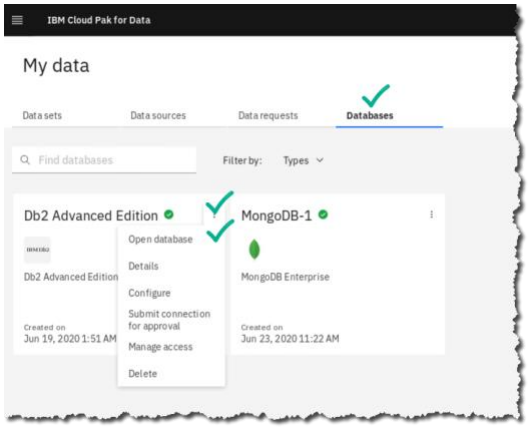
3.3.2 Reviewing the Db2 Advanced Edition data

__4. To review this data, you will be launching the *Db2 Advanced Edition Data Management Console*. In the CPD web client, click the `Navigation Menu` (“hamburger” icon).

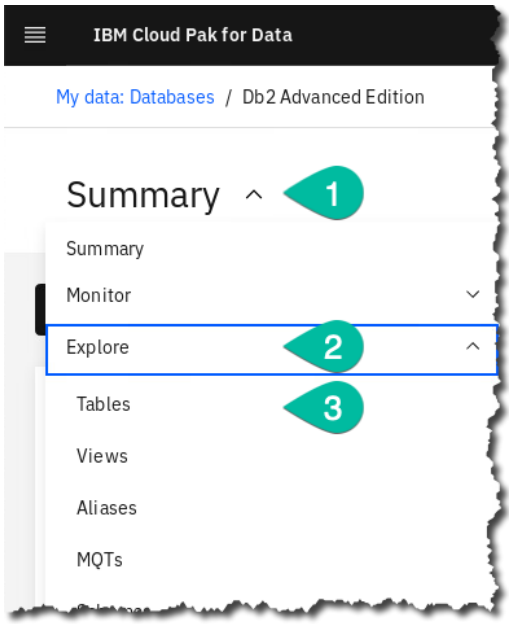


__5. Click `Collect` ⇨ `My Data` ⇨ `Databases` ⇨ `Db2 Advanced Edition` ⇨ `ellipsis` ⋮ ⇨ `Open database` *(Note: Please wait for the green check to appear next to Db2 Advanced Edition first.)*

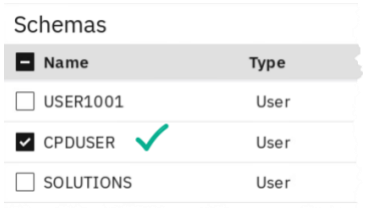




__6. At the top left of the Db2 Advanced Edition console, click on: `Summary` ⇨ `Explore` ⇨ `Tables.`

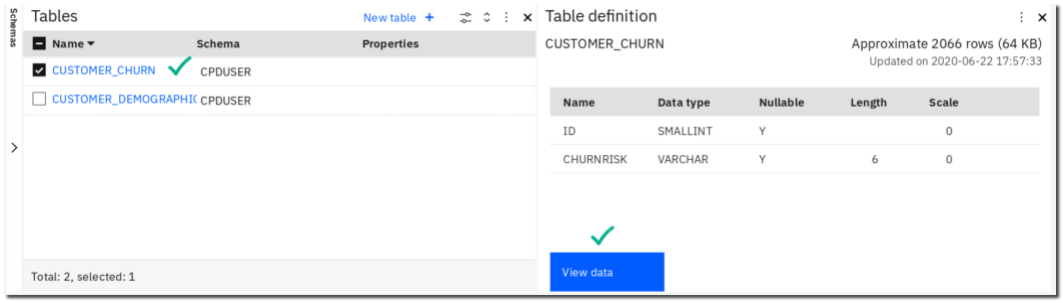


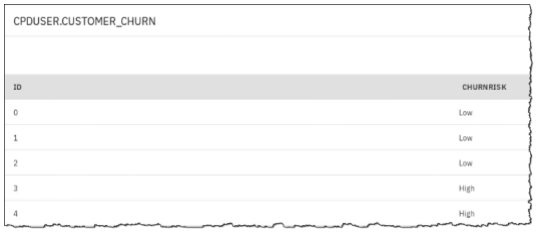
__7. Select schema `CPDUSER.`



__8. Click `CUSTOMER_CHURN` and click on it to bring up the table definitions view.

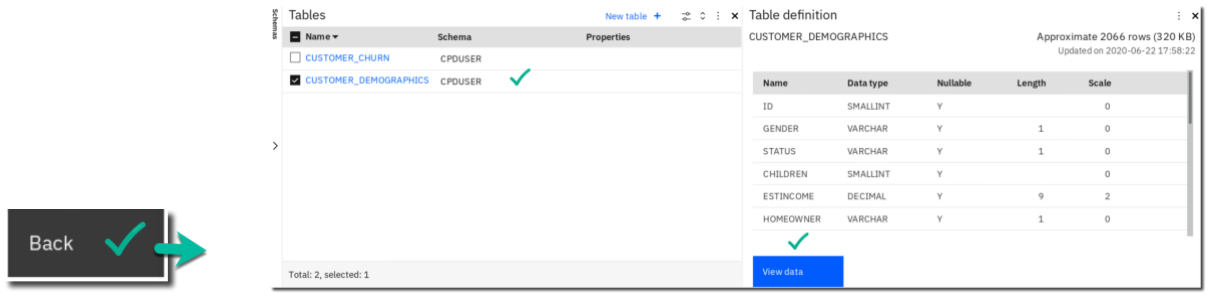
Click on `View Data.`

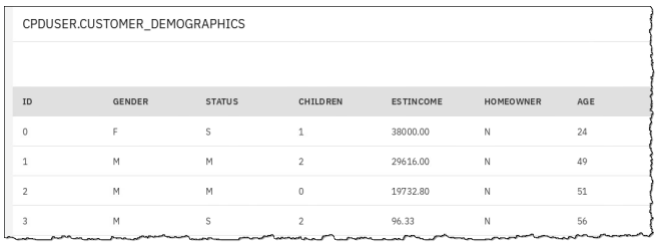




This data represents churn risk that was determined through a manual methodology by Trade Co. It will be enhanced by a machine learning methodology created in this workshop.

__9. Click on the `Back` icon, then table `CUSTOMER_DEMOGRAPHICS` ⇨ `view Data.`



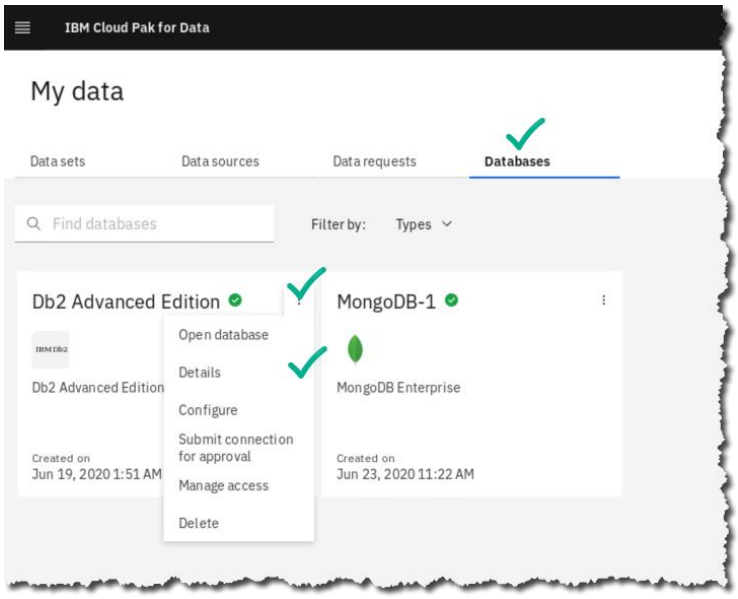


This data represents the known characteristics of the Trade Co. customers themselves. The machine learning methodology of this workshop will take advantage of this data to determine churn risk more accurately.

3.3.3 Reviewing the Db2 connection

__10. In order to create a connection to a CPD data source, you would first need the connection information. To review what that looks like for the in-cluster Db2 Advanced Edition, click:

`Navigation Menu` ⇨ `Collect` ⇨ `My Data` ⇨ `Databases` ⇨ `Db2 Advanced Edition` ⇨ `ellipsis ⋮` ⇨ `Details`



__11. This takes you to the details page for this data source.

On the right you will find the `Access Information` section which gives you the information you need to create a connection.

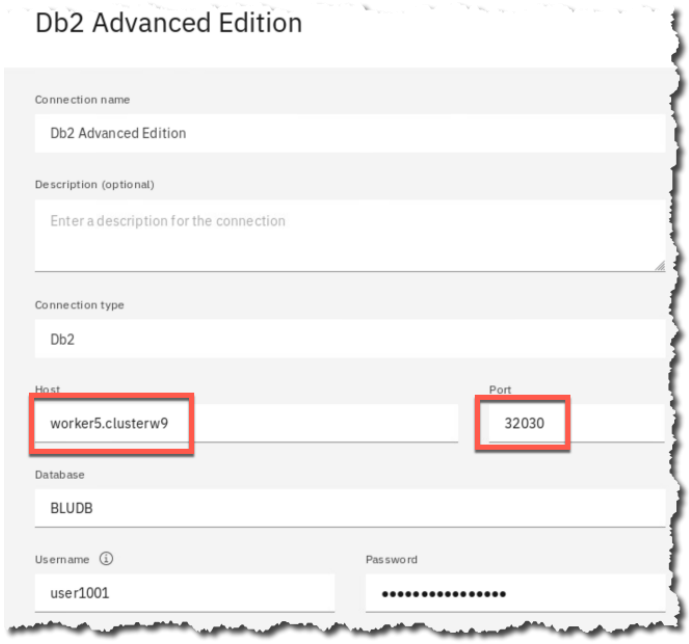
Note: `Username` and `Password` are explicitly shown (and easily copied using the icons) while the JDBC Connection URL string tells us the `Host`:`Port`/`Database.`

__12. Now you can review the connection to the Db2 Advanced Edition that was already created for this workshop.

Go to: `Navigation Menu` ⇨ `Connections` ⇨ `Db2 Advanced Edition`

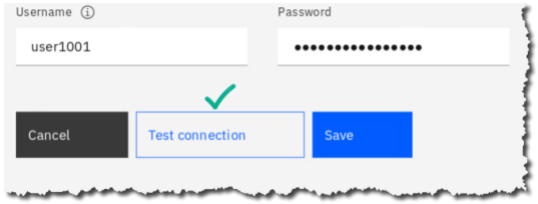
__13. Hover over the left part of the connection and click the `Edit` (pencil) icon.

__14. The Db2 Advanced Edition `Connection` edit screen looks like this:



Each data source type (Db2, Mongo, Hive, S3, Drop Box, and so on) has their own connection page format. Many of the relational database sources are similar to this one, while others require different credentials altogether.

__15. At the bottom of the screen, click `Test connection.`



__16. At the top left of the screen, you should see this Success message.

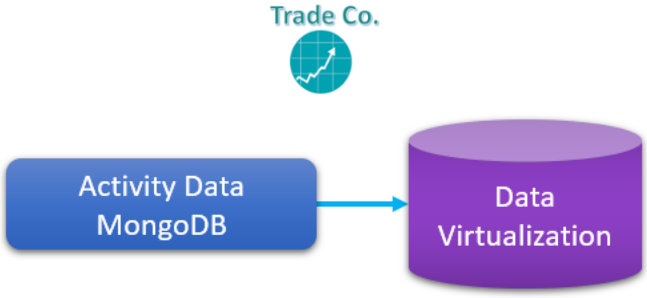
__17. To exit the edit screen, click `Cancel.` DO NOT SAVE this connection.

 Data Engineer	Note: if you do NOT have a successful connection test as shown above, you should fix this by launching another CPD web client from the desktop and use the second web client to return to the Details screen of the Db2 Advanced Edition. Use both screens to make sure all the credentials match between the Details screen and the Connection screen, especially Username/Password.

3.4 MongoDB data overview – Virtualizing for analytics

The Mongo data in our scenario (Activity) has two key factors that makes for an appropriate scenario for Data Virtualization:

- The data is constantly changing (it is dynamic)
- Our analytics processing requires the latest data



We have chosen semi-structured data in this scenario to demonstrate the power of CPD Data Virtualization – it even works with JSON documents. But we could just have easily chosen a structured database source for it, like Db2, Informix, SQL Server, and so on.

Semi-structured (and even unstructured) data is commonly used in many systems of engagement applications (like Trade Co.'s Stock Trader) so this was another reason we chose this as an example data source type.

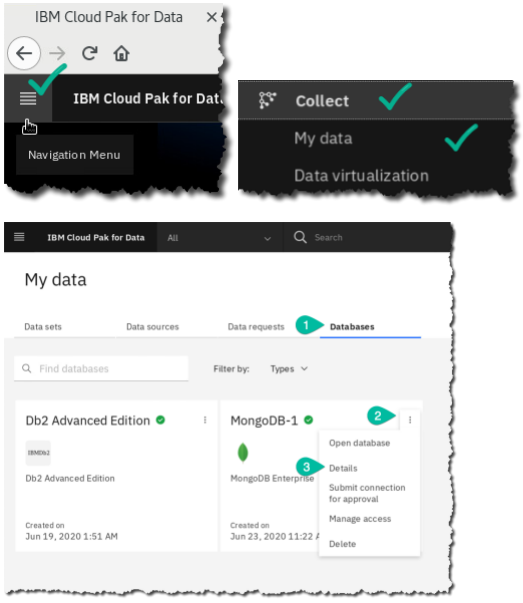
Our scenario presumes that this data comes from a mobile application that will be constantly changing the data, and that we would require the latest data for our analytics workflow.

3.4.1 Reviewing the MongoDB data

The MongoDB database was provisioned after the MongoDB Enterprise service was installed, which is also located in `My data`.

__18. Start at the `Navigation Menu.`

__19. Click `Collect` ⇨ `My data` ⇨ `Databases` ⇨ `MongoDB-1` ⇨ `ellipsis` ⋮ ⇨ `Details.`

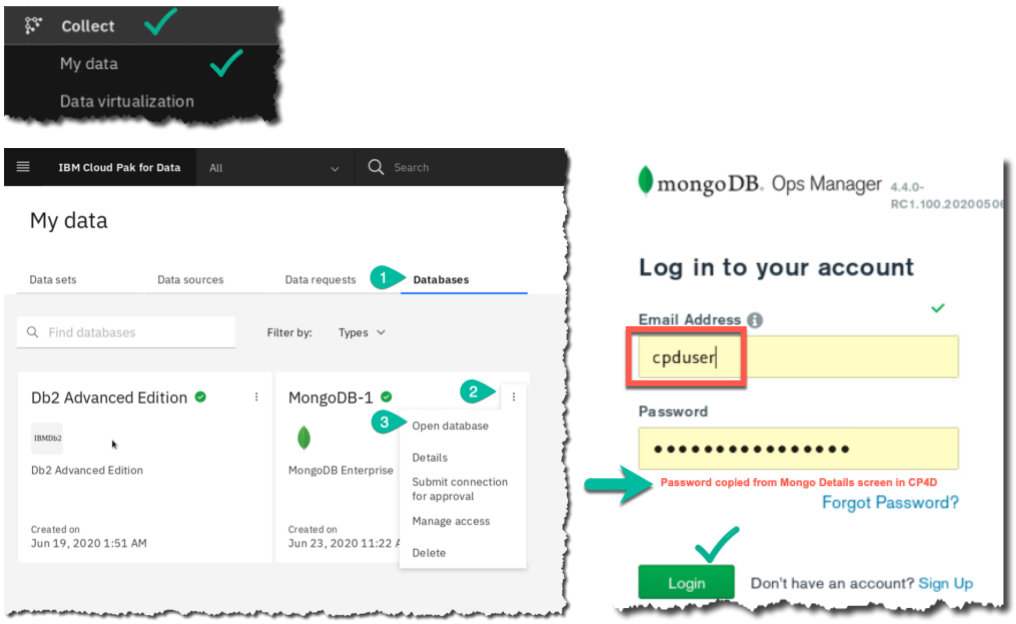


__20. Examine the section `MongoDB Ops Manager` on the right.

Here you can review the username for the Ops Manager (which is cpduser).

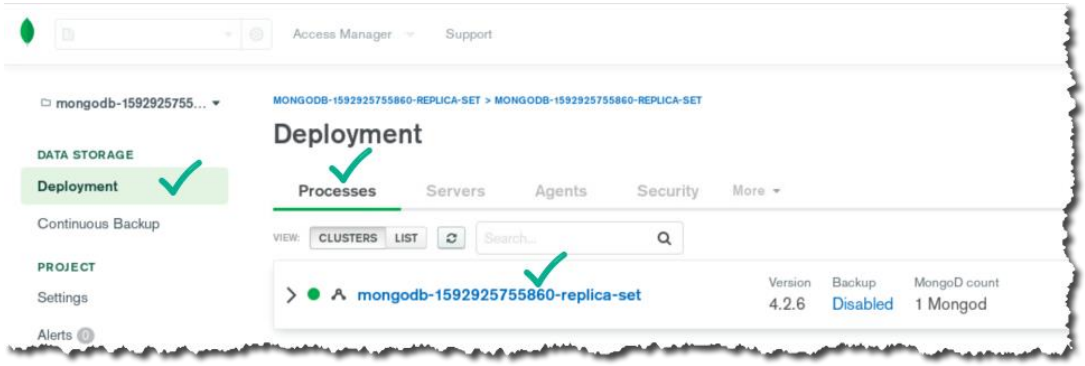
Copy the `Password` using the icon.

__21. Now return to the `My data` screen and choose `Databases` ⇨ `MongoDB-1` ⇨ `ellipsis` ⋮ ⇨ `Open Database.`

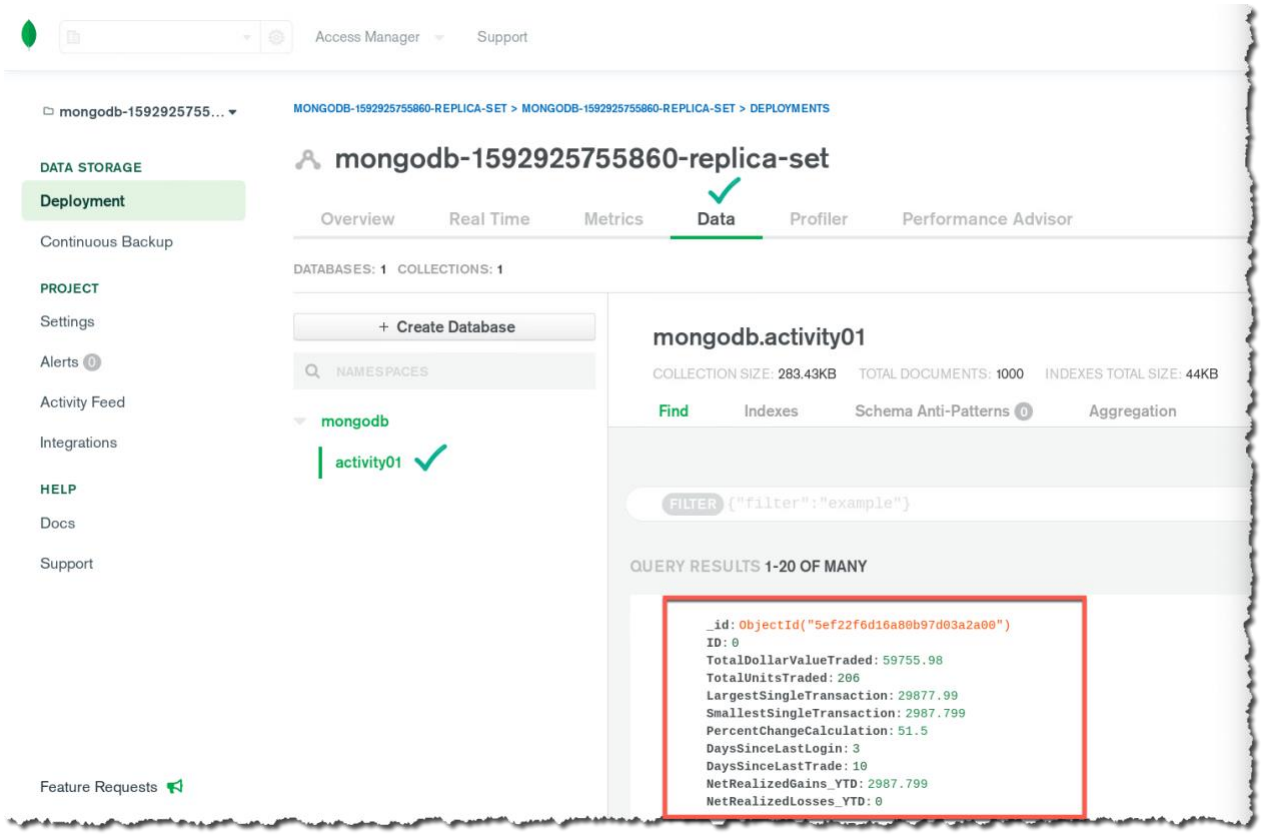


Data Engineer	The very first time you log into this *MongoDB Ops Manager* console it may not ask you for user and password credentials. We have saved these for you in a browser setting. However, if the credentials were not saved and should you encounter this Login screen (as above) you can use the password copied from the `Details` screen. Note: if the copy/paste of the password does not work for you using `Details` screen icon, then try this technique: 1. In the `Details` screen, perform the "copy" by highlighting the password and `right-click` ⇨ `copy`. 2. In the `Ops Manager` login screen use `right-click` ⇨ `paste`.

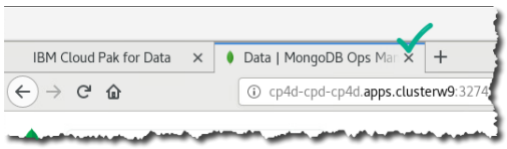
__22. In the *MongoDB Ops Manager* console tab, click `Deployment` on the top left of the screen. In the `Processes` tab, click on the `Replica Set` link,



__23. Click on the `Data` tab to review the JSON documents in database `mongodb.activity01.`

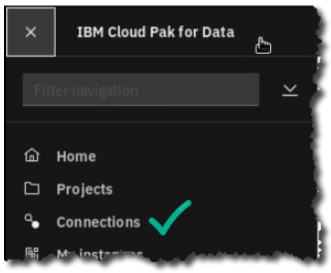


__24. When finished, close the tab `Data | MongoDB Ops Manager.`



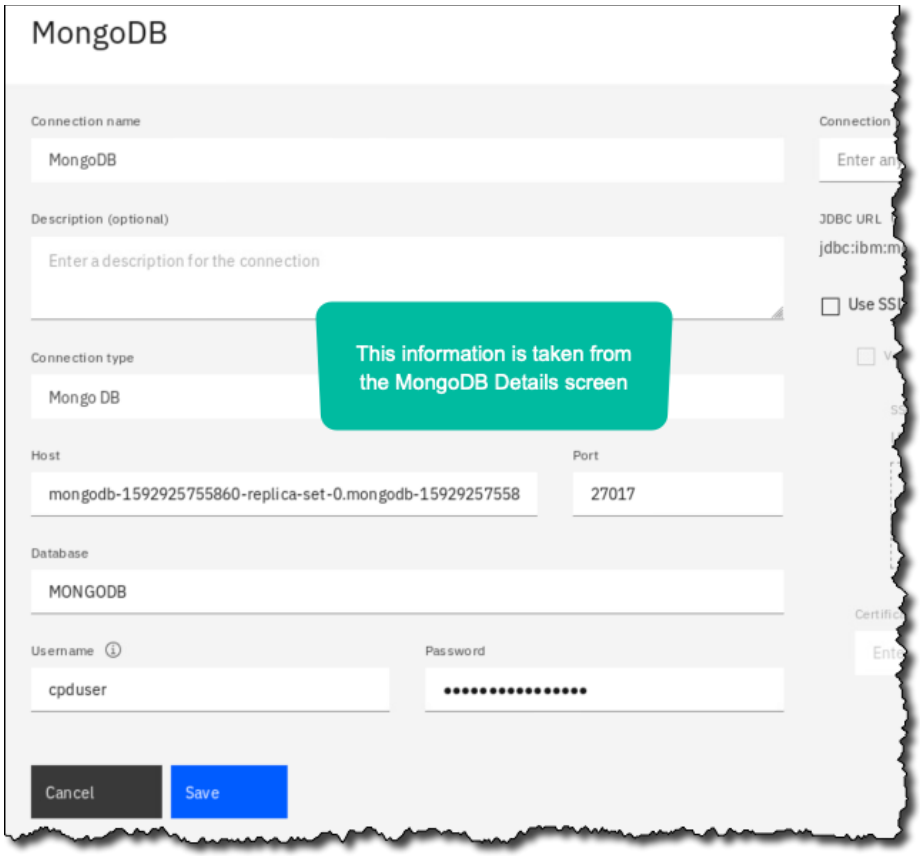
3.4.2 Reviewing the MongoDB connection

__25. Go to `Navigation Menu` ⇨ `Connections`



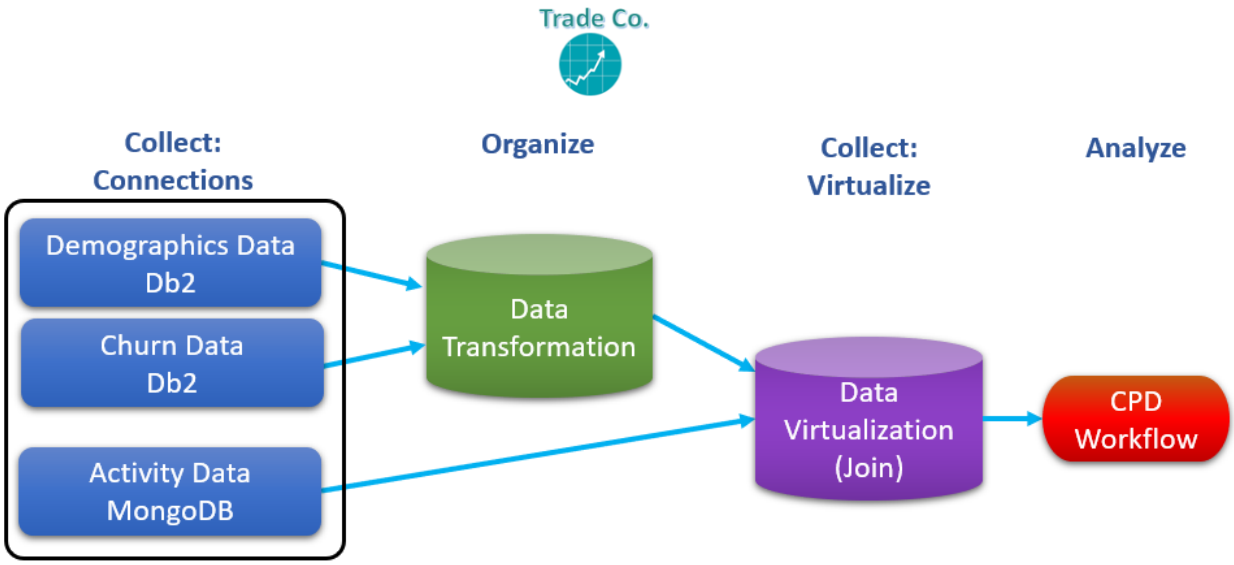
__26. Click on connection `MongoDB` line and choose the `Edit` icon.

__27. Note that the credentials information is taken from the Details screen you reviewed earlier.



3.5 Lab conclusion

In this Collect: Connections lab, you reviewed connections for the Db2 Advanced Edition and MongoDB data. This prepares you for the Organize and Collect: Virtualize labs in this workshop.



In the `Organize` lab coming up, we will be transforming the Db2 data sets into one.

After that, we will finish up the `Collect` processing by virtualizing the results from both the Db2 Transformation output and the MongoDB data together.

**** End of Lab 03 - Collect: Connections**

Lab by Burt Vialpando and Kent Rubin, IBM

Lab 04 ORGANIZE

4.1 Lab overview

Many organizations find it difficult to understand their own data because it originates from many sources, is dispersed across many silos, and is controlled by different teams.

This `Organize` lab will show you how to uncover the hidden value in your organization's data and how to build a lineage that is otherwise difficult to establish. Cloud Pak for Data helps your organization move from the manual processes required to establish relationships between data to an automated one aided by the platform's built-in machine learning capabilities.

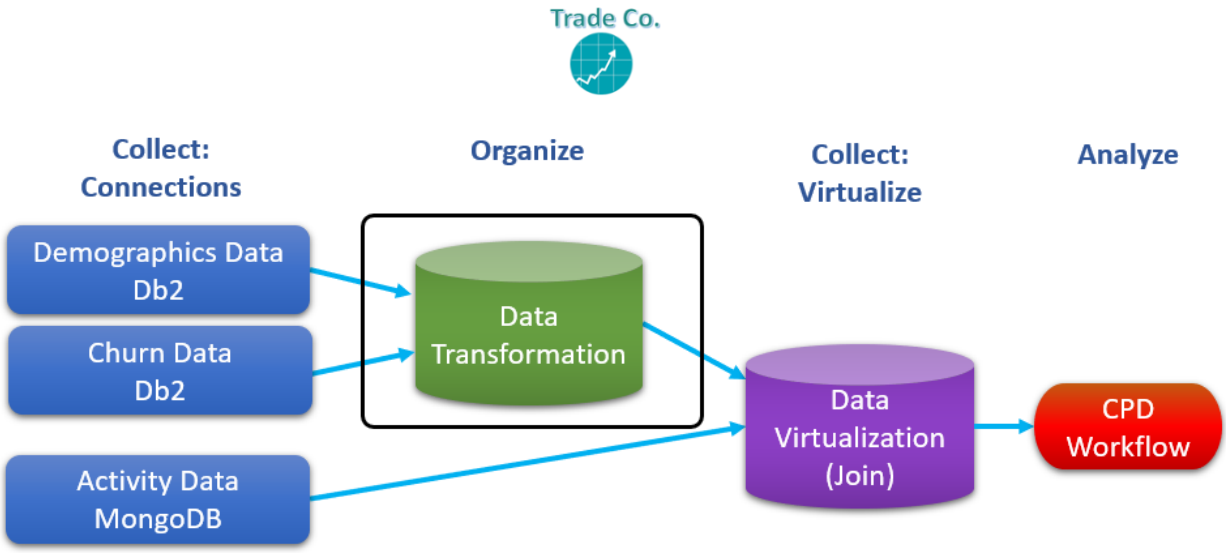
4.2 Persona represented in this lab

The `Data Steward` persona is the likely one to perform most of the `Organize` tasks shown in this lab.

Persona (Role)	Capabilities
 Data Steward	Data Stewards integrate and transform data as well as provide governance, lineage and classification of the data.

Note: this persona often works closely with the `Data Engineer` because they both work with the data to prepare it for analytics processing by other personas. For example, in this lab one of activities the `Data Engineer` will do is to create a Transformation job with the Db2 data. In the next lab, the `Data Engineer` will go on to use that output table to create a final virtualized view of all the data sources joined together.

Note: The Data Steward persona also works closely with the `Data Quality Analyst` persona.



Before we start transforming data, let's first explore the other crucial aspects of the CPD `Organize` capabilities.

4.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Double-click the desktop icon: `Cloud Pak for Data Web Client.`

__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

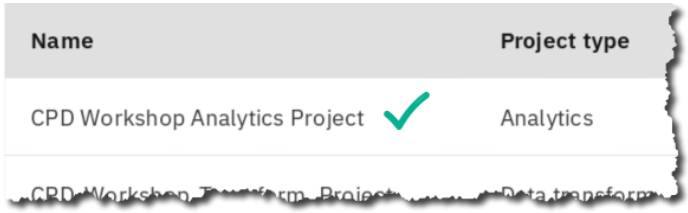
4.4 Reviewing a data asset in the project

Some of the more interesting “Organize” capabilities can be done on a data asset that has been added to a project. In this Trade Co. scenario, the team has created a Project from which they will all collaborate and work from together.

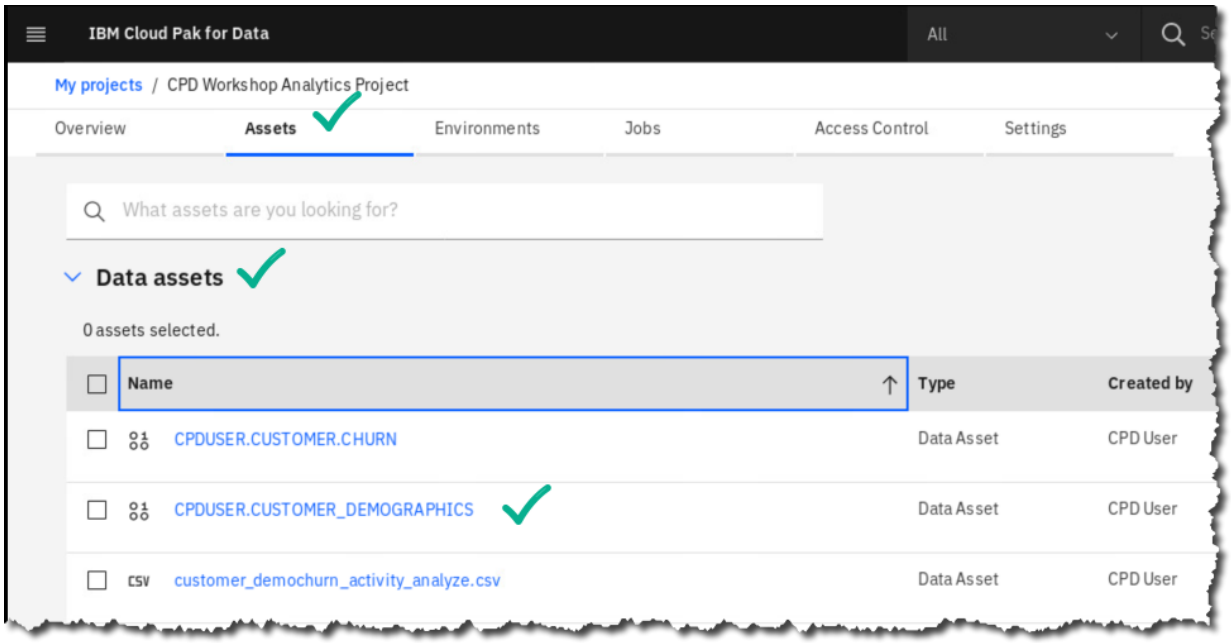
4.4.1 Data asset overview

__4. In the CPD web client, click the `Navigation Menu` (“hamburger” icon) ⇨ `Projects.`

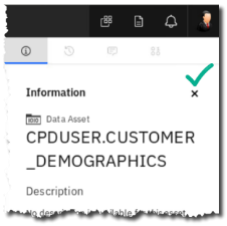
__5. Select the project: `CPD Workshop Analytics Project`.



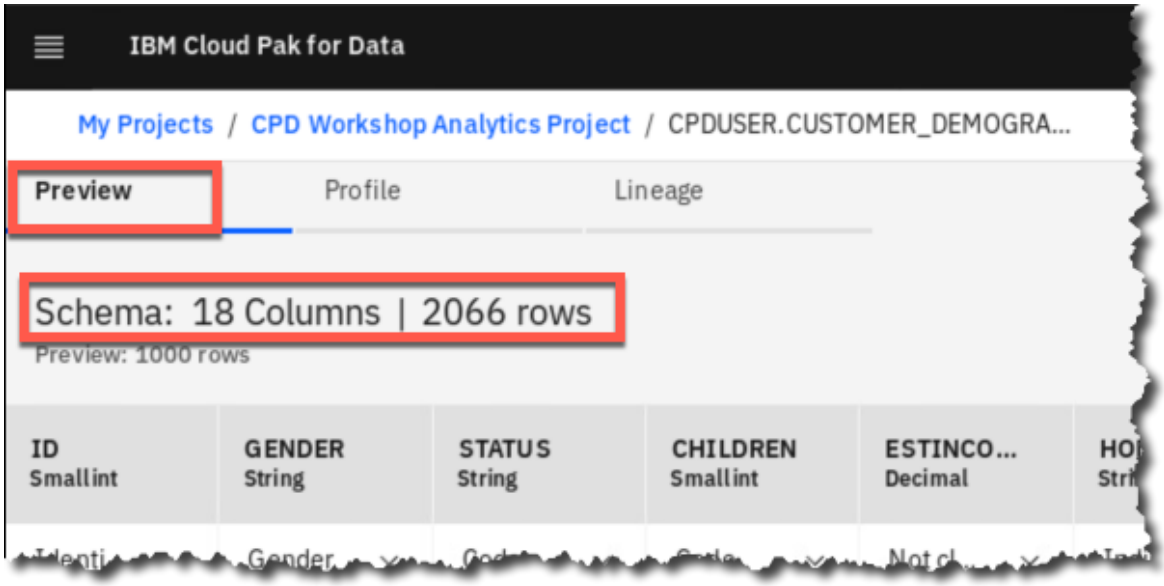
__6. Under tab `Assets` ⇨ `Data assets,` click `CPDUSER.CUSTOMER.DEMOGRAPHICS.`



__7. Close (`x`) the Information window on the top left.

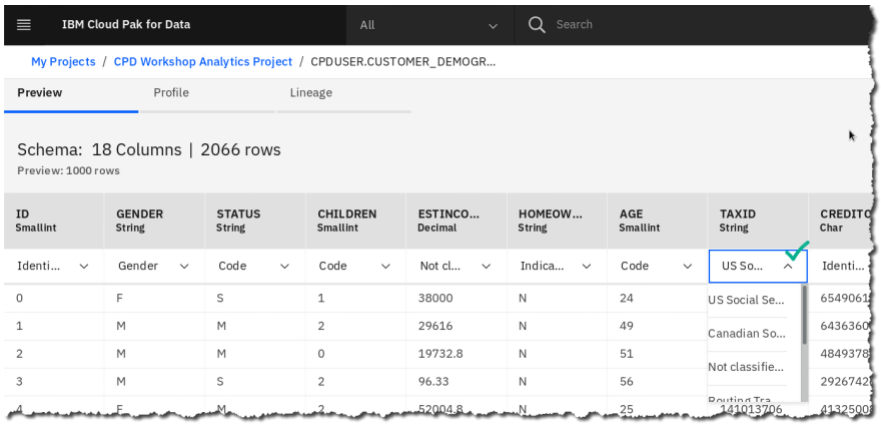


__8. In the `Preview` section, general information about the data asset is displayed, as well as column specific information and sample data.



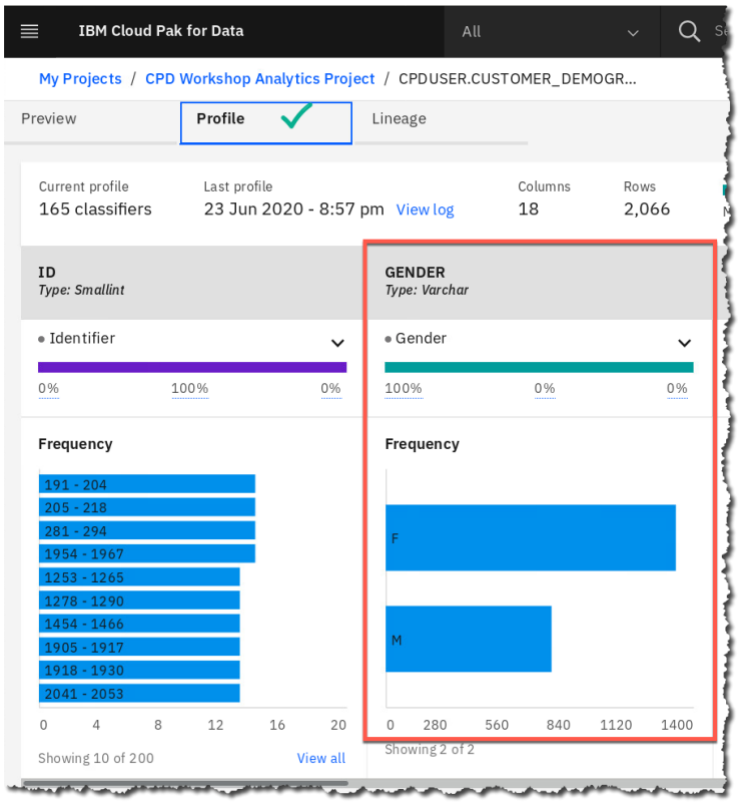
__9. Click on the down arrow for column TAXID to see how Profiling categorized the data.

The identifier is: US Social Security Number.



__10. Click Profile to see more details on the profiling information. (It may take a minute or two to render – be patient.)

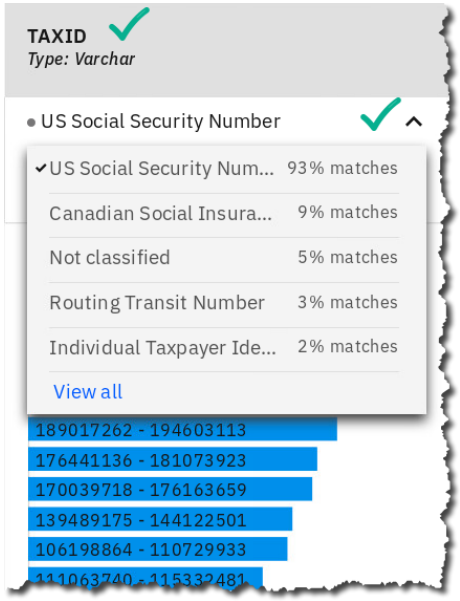
Notice column GENDER was assigned Identifier Gender, which has 2 of 2 values (M and F) and a displayed frequency for each.



Data Quality Analyst

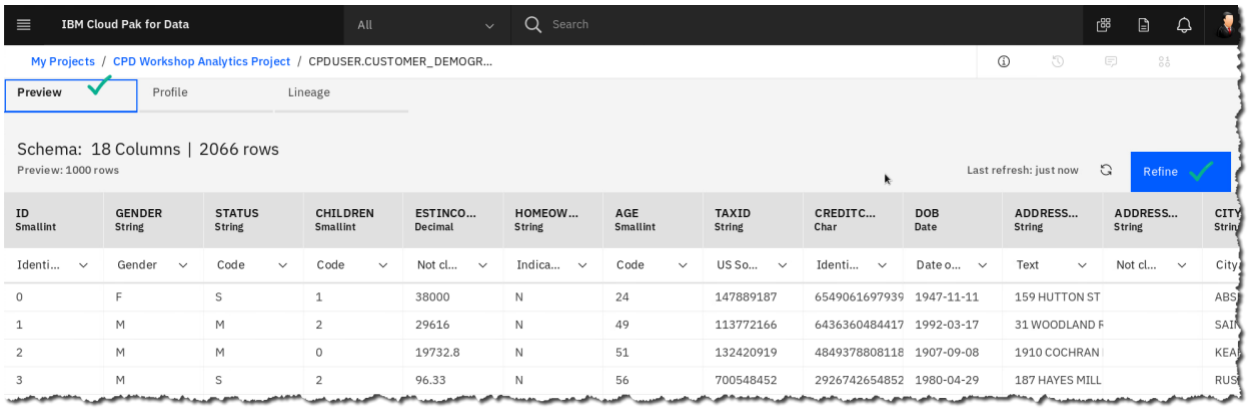
Note: This profiling was performed by the Data Quality Analyst persona

__11. Scroll over to column `TAXID` and click the twisty to see how this data is profiled.



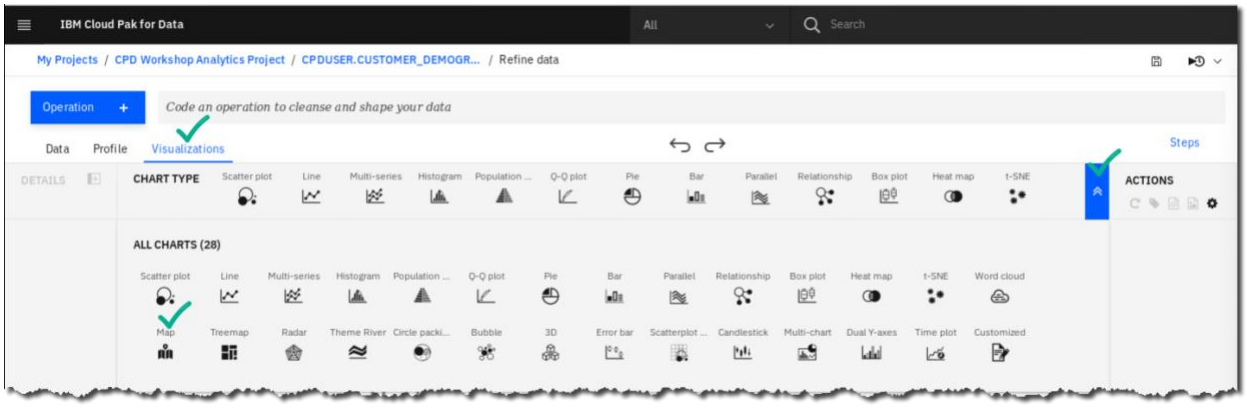
4.4.2 Refine visualizations

__12. Click back to `Preview` then click `Refine`

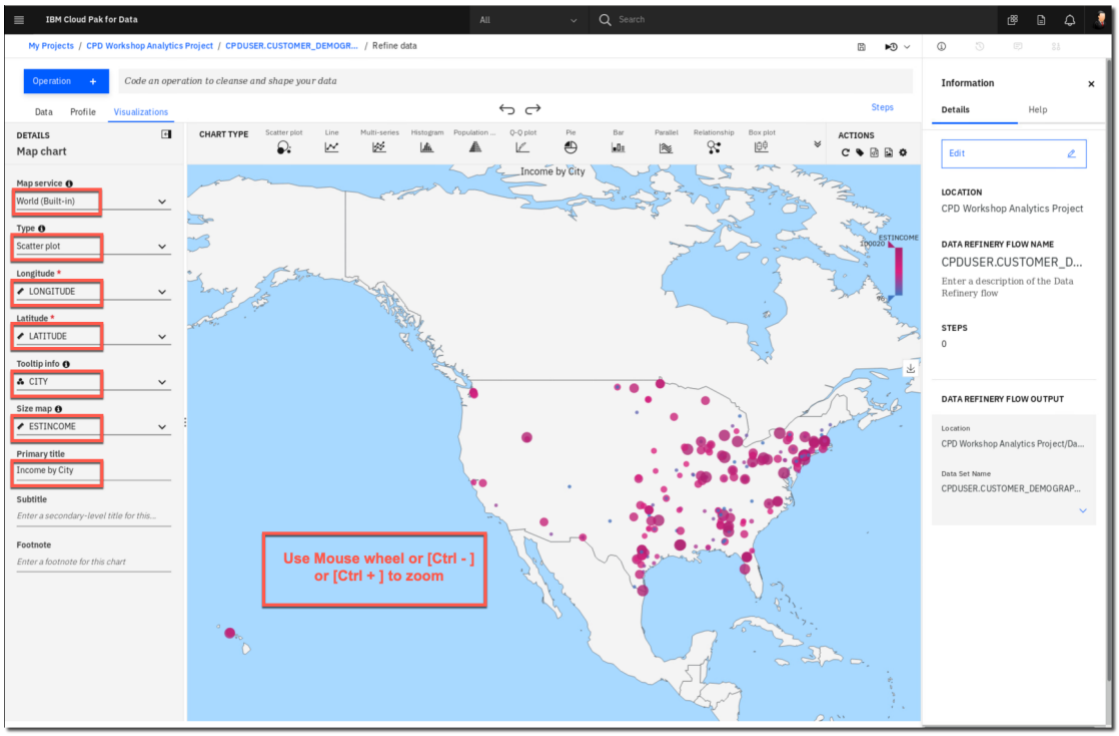


__13. Click `Visualizations` ⇨ `Map`

(Hint: you may have to select the double down arrow to see `Map`.)

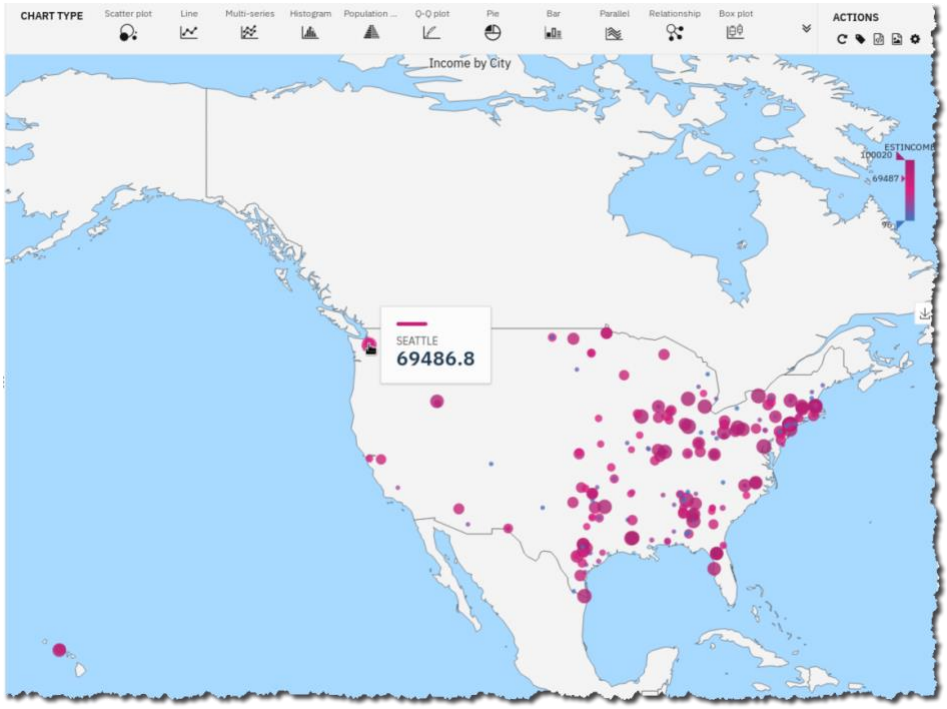


__14. Fill in the Details as World, Scatter plot, LONGITUDE, LATITUDE, CITY,ESTINCOME and Income by City



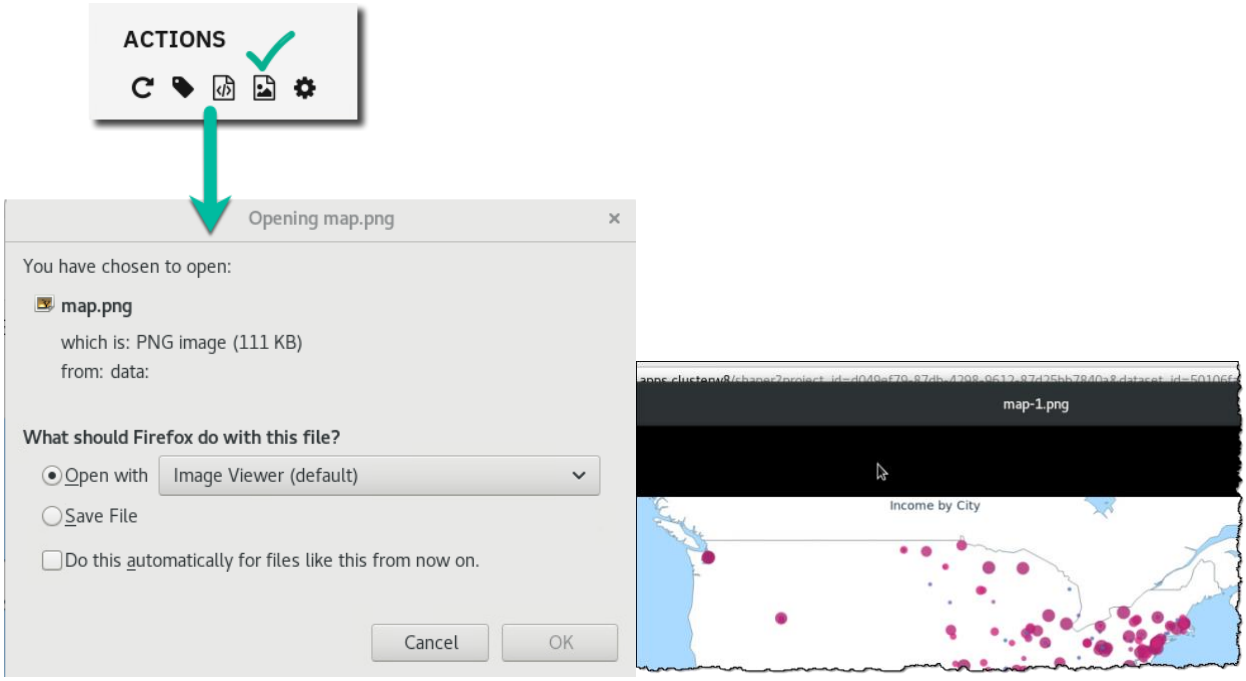
__15. Use your mouse's wheel to Scroll up and center the United States on the visualization. Use the [Ctrl] – and [Ctrl] + to zoom in and out to properly size the visualization.

__16. Hover over any circle on the map to see the Estimated Income for that City.

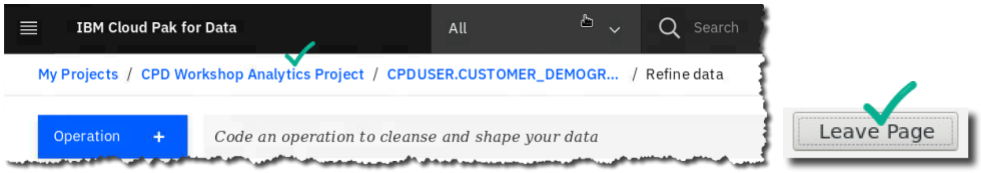


__17. If the Data Scientist or Business Analyst wants to keep a visualization, they can download it by clicking on the `image icon` shown.

They can save or review it by downloading it from the `ACTIONS` section and then using `Open with Image Viewer.`



__18. Close the Image viewer (if you have it open) and click on the breadcrumb back to the link for the `CPD Workshop Analytics project` ⇨ then `Leave Page.`

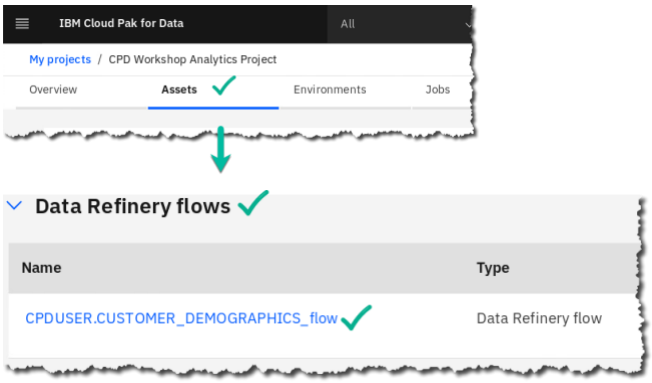


4.4.3 Data refinery flows

You should be back at the project `CPD Workshop Analytics Project`. If you are not there, navigate there as you did at the beginning of this lab.

__19. Under `Assets`, scroll down to find `Data Refinery flows`

Click `CPDUSER.CUSTOMER.DEMOGRAPHICS_flow`

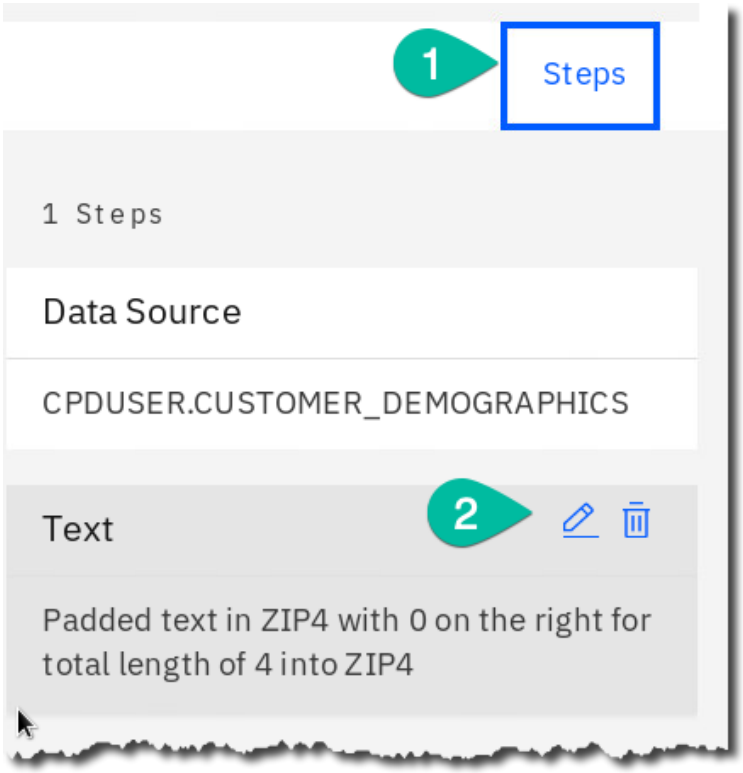


__20. This Refinery flow was pre-created for you to shorten this lab. It was created in the Refine screen you were in earlier. Click the `Steps` Box (1).

There is only one step in this job, which is a “Text” operation on the ZIP4 column.

`Hover` over the step to see the icons for it.

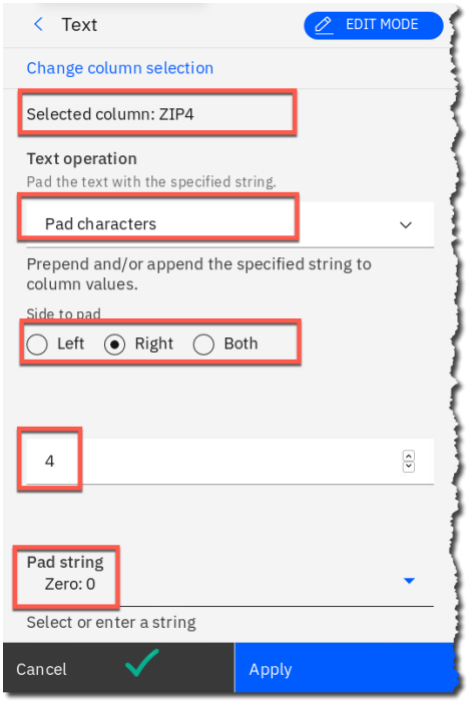
Click the `edit` (pencil) icon (2) to review the details of the step.



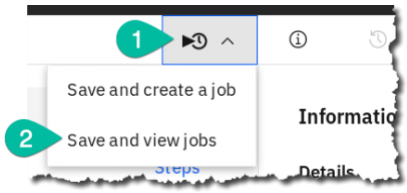
__21. This step refines the ZIP4 column by padding 4 characters of “0” to the Right.

This means if the column is already filled in, it will not do anything. If it is empty or filled with 0, it will fill it in with 0000.

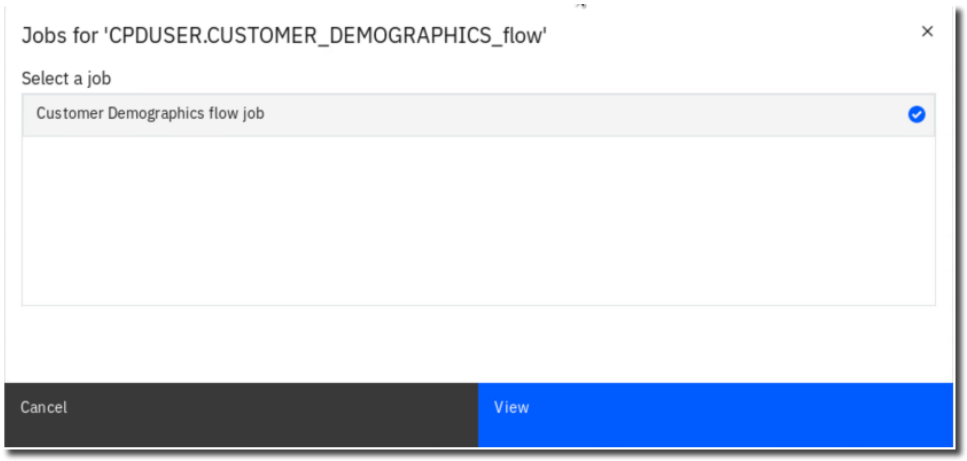
Click `Cancel` (the job is OK as it is, but you could have changed it here if you needed to).



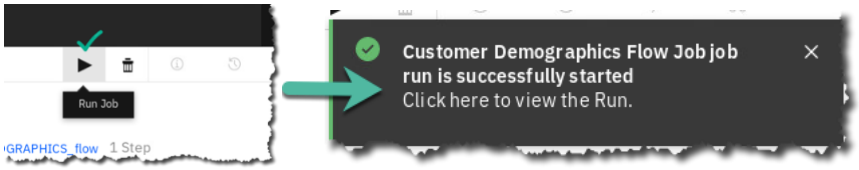
__22. Click on the `Jobs` (arrow by a clock) `icon`, then `Save and view jobs.`



__23. Click the `Customer Demographics Flow Job` shown, then `View.`



__24. Click the `Run Job` icon.



Note: If the job fails for any reason, reselect the source csv file to prime it and rerun the job.

__25. The job takes about a minute to complete. You can see it Running on the screen.

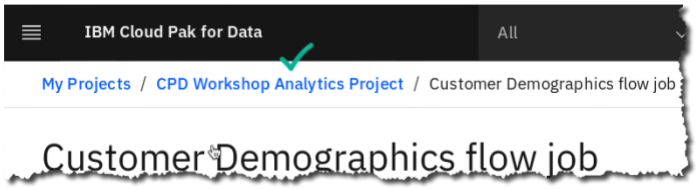
Hit the `refresh` icon to see status more quickly.

__26. When `Completed`, you can check the logs of the job if you want to. This can be especially useful if the job is complex.

Click on it to review the details of the job.

Data Steward	If the job run fails, simply click the ellipses on that job run and delete it, then run it again.

__27. Click on the breadcrumb trail to return back to the project.

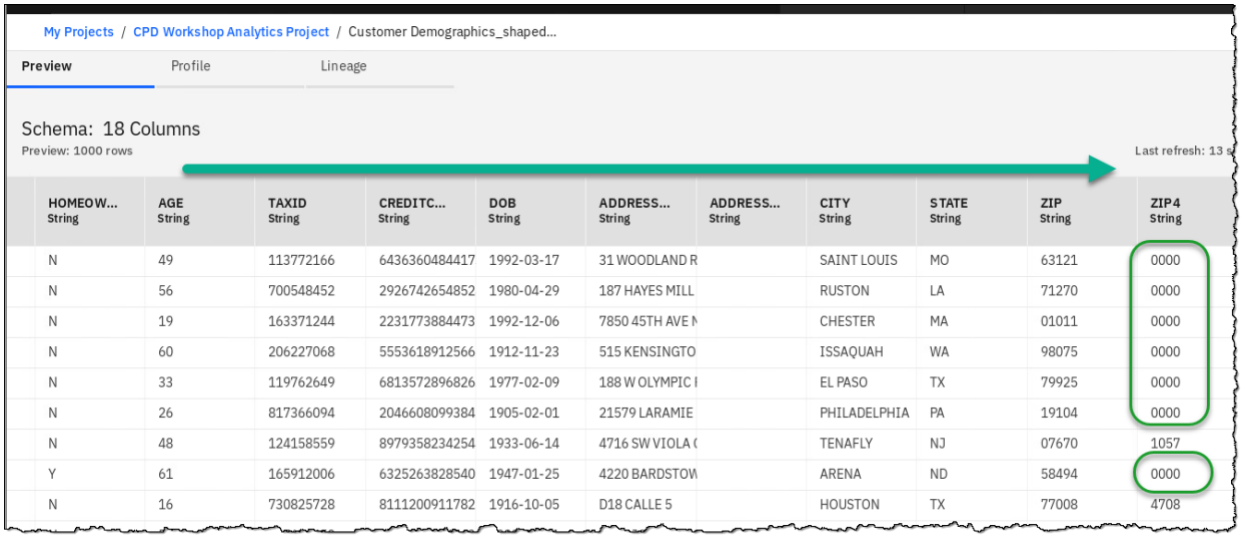


__28. A new `Data Asset` was created as output from the Flow (refine job).

Scroll down to find and then click on `Customer Demographics_shaped.csv`

__29. Scroll over to find the ZIP4 column.

Notice the data is padded with four zeroes if it was less than four zeroes. If the row already had a four-digit value, it was left alone.



HOMEOW... String	AGE String	TAXID String	CREDITC... String	DOB String	ADDRESS... String	ADDRESS... String	CITY String	STATE String	ZIP String	ZIP4 String
N	49	113772166	6436360484417	1992-03-17	31 WOODLAND R		SAINT LOUIS	MO	63121	0000
N	56	700548452	2926742654852	1980-04-29	187 HAYES MILL		RUSTON	LA	71270	0000
N	19	163371244	2231773884473	1992-12-06	7850 45TH AVE N		CHESTER	MA	01011	0000
N	60	206227068	5553618912566	1912-11-23	515 KENSINGTO		ISSAQUAH	WA	98075	0000
N	33	119762649	6813572896826	1977-02-09	188 W OLYMPIC I		EL PASO	TX	79925	0000
N	26	817366094	2046608099384	1905-02-01	21579 LARAMIE		PHILADELPHIA	PA	19104	0000
N	48	124158559	8979358234254	1933-06-14	4716 SW VIOLA (		TENAFLY	NJ	07670	1057
Y	61	165912006	6325263828540	1947-01-25	4220 BARDSTOW		ARENA	ND	58494	0000
N	16	730825728	8111200911782	1916-10-05	D18 CALLE 5		HOUSTON	TX	77008	4708

Data Steward

The power of `Refine` can be used by more than just the Data Steward because it is launched from a Project, which enables many more personas the ability to shape and refine data assets to which the Data Steward may have given them access. This gives anyone the ability to explore and shape data on their own with self-service capabilities.

If the need is to make a permanent change to the data, for example, to write it back to a database from where it came, that is where `Transform` comes in. We will explore Transform later in this lab.

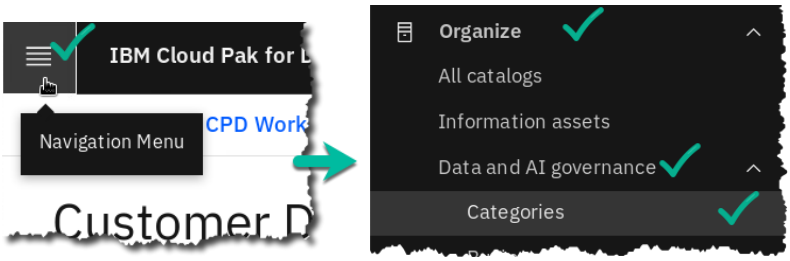
4.5 Reviewing a business glossary

A business glossary consists of categories and terms.

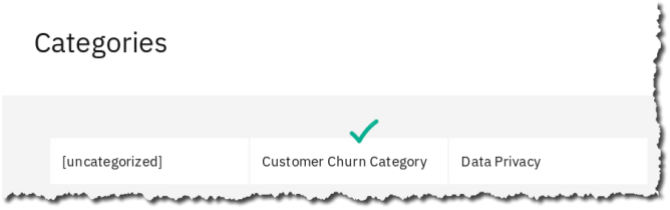
Categories provide the logical structure for the glossary so that you can browse and understand the relationships among terms and categories in the glossary. Categories can be organized in a hierarchy based on their meaning and relationships to one another.

A Business term is a word or phrase that describes a characteristic of the enterprise. Business terms are the fundamental building blocks of the glossary. Each Business term has a parent Category, but it can also be referenced by other Categories. When you create a Business term, you need to provide a meaningful name. Business terms can be assigned to other Business terms, and to other asset types as well.

__30. Click Navigation Menu ⇨ Organize ⇨ Data and AI governance ⇨ Categories.



__31. Review a category already created for you: Customer Churn Category.

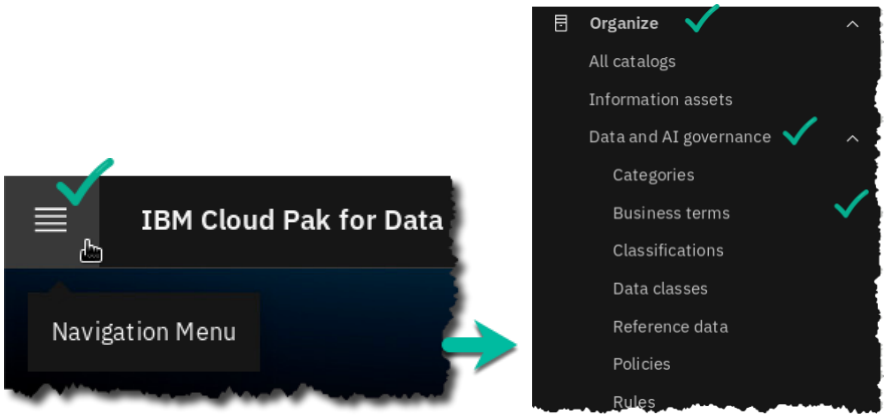


__32. Review the Business terms and Policy for the Category.

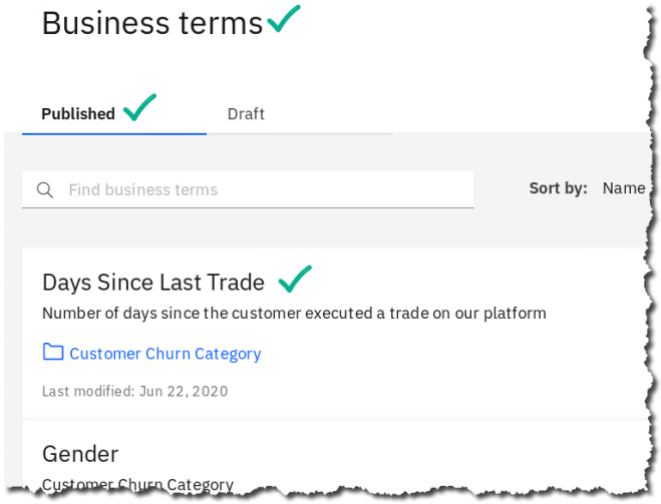
Name	Description	Type
Days Since Last Trade	Number of days since the customer executed a trade on our platform	Business term
Gender	Customer Churn Category	Business term
Home Owner	Flag indicating whether a customer owns a home	Business term
Income	Annual household income of the customer	Business term
Net Gains and Net Losses are Mutually Exclusive	A customer can only have a value in Net Gains or Net Losses	Policy

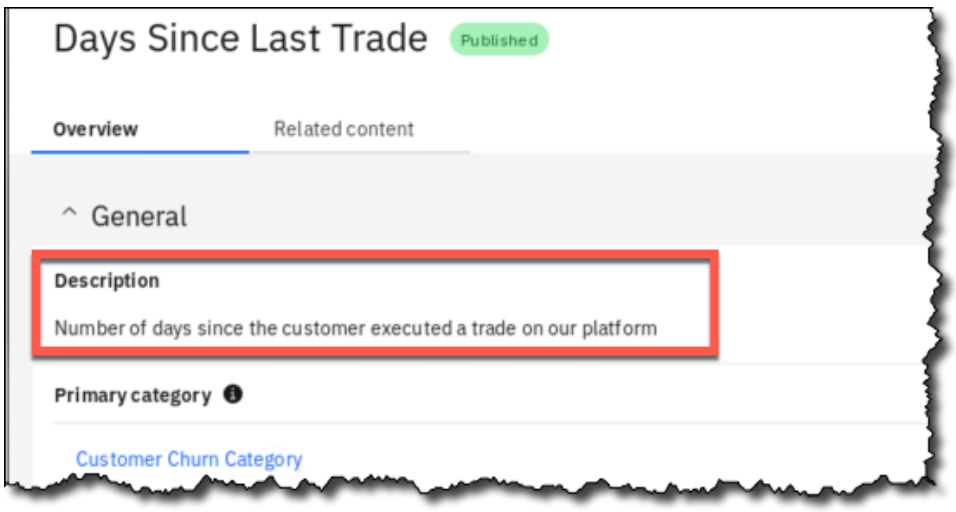
Note: You can view more than 5 items per page.

__33. Click `Navigation Menu` ⇨ `Organize` ⇨ `Data and AI governance` ⇨ `Business terms.`



__34. Here you can edit or add new Business terms that are either in Published or Draft mode. Click on one to review it in more detail.





Data Steward

You can create your own Glossary with Categories and Business terms manually, or import them from a file. This workshop was prepared using .csv files in in the following directory on your OS:

```
[root@master1 organize]# pwd
/workshop/labs/organize
[root@master1 organize]# ls *.csv
-rw-rw-r--. 1 ibmdemo ibmdemo  980 Aug  4 16:39 organize-businessterms.csv
-rw-rw-r--. 1 ibmdemo ibmdemo  845 Aug  4 16:39 organize-businessterms-fixed.csv
-rw-rw-r--. 1 ibmdemo ibmdemo  264 Aug  4 16:39 organize-categories.csv
-rw-rw-r--. 1 ibmdemo ibmdemo  329 Aug  4 16:39 organize-policies.csv
-rw-rw-r--. 1 ibmdemo ibmdemo 1245 Aug  4 16:39 organize-referencedatasets.csv
-rw-rw-r--. 1 ibmdemo ibmdemo 2438 Aug  4 16:39 organize-rules.csv
[root@master1 organize]#
```

In addition, you can import Industry Models from IBM for industries such as finance, banking, healthcare, and insurance and import them into CPD.

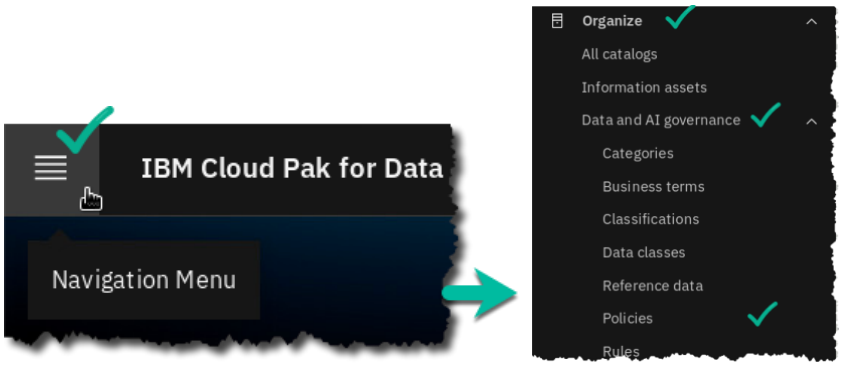
See the services screen  then `Industry Accelerators.`

4.6 Reviewing a Governance Policy and a Rule

An information governance `Policy` is a natural-language description of a governance subject area. It can contain multiple information governance sub-policies or reference one or more information Governance Rules. It must fulfill a business objective and be relevant and understandable to all users of the policy.

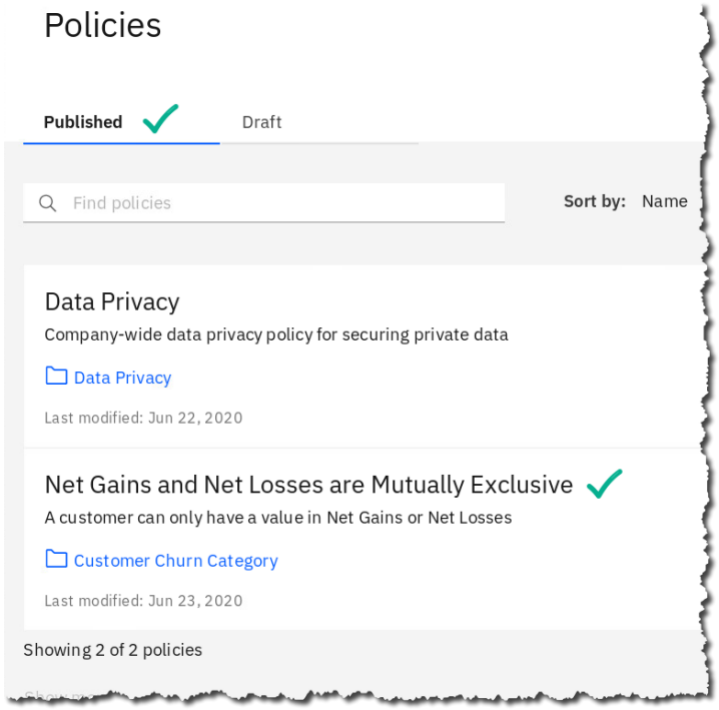
An information Governance `Rule` is a natural-language description of the criteria that are used to determine whether information assets are compliant with business objectives. Generally, information governance rules are derived from information governance policies and are more specific. The Rules define the actions to take in specific situations to implement the Policy. An information Governance Rule can be referenced by one or more information Governance Policies.

__35. Click `Navigation Menu` ⇨ `Organize` ⇨ `Data and AI governance` ⇨ `Policies.`



Note: the first time using the page may take a minute or two to render – be please patient.

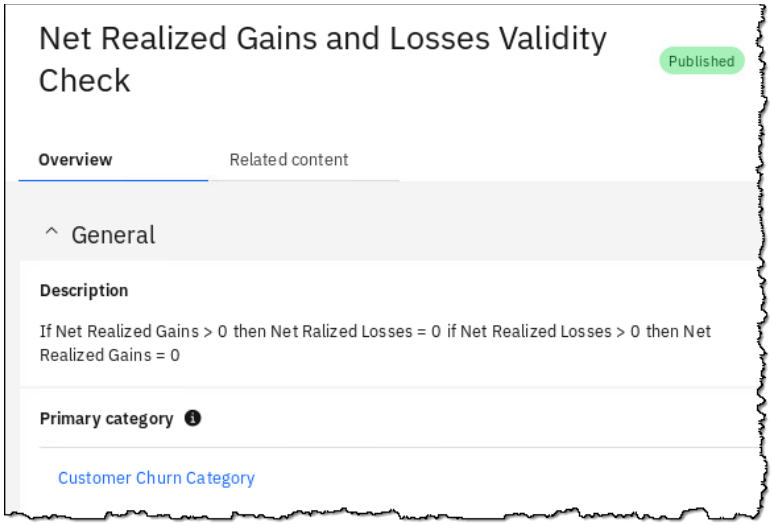
__36. Review the Published Policy that has been associated with our previously reviewed Category.



__37. Scroll down until you find the Rule that is associated with this Policy.

Notice you could add another rule to this policy here. (Don't do this now.)

Review the Rule by clicking on it.



Data Steward

A `Data Dictionary` contains a Business Glossary (Categories and Business terms) as well as information Governance Policies and Rules to ensure data compliance with business objectives.

In this lab we have a beginning sample of these items, but in reality, a Data Dictionary for any organization is quite large and can and should be updated as frequently as new data sources, regulations, and other criteria require it.

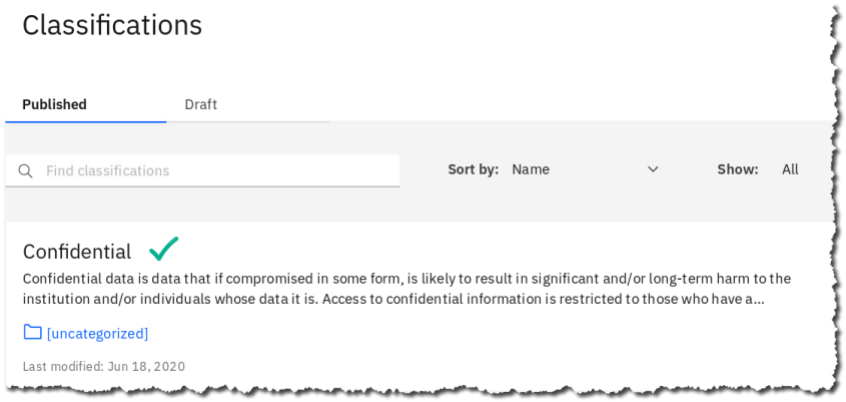
4.7 Reviewing Classifications, Data Classes, and Reference Data

4.7.1 Classifications

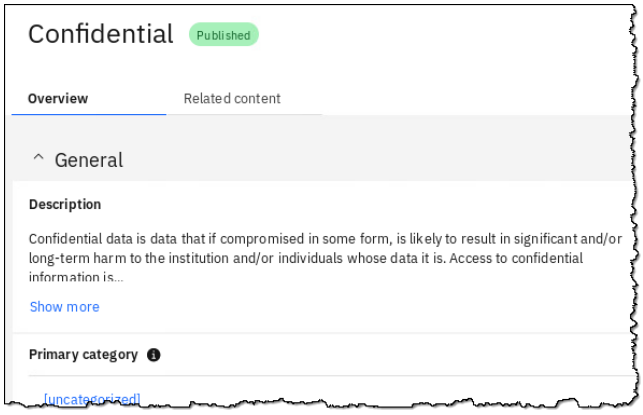
A `Classification` describes the sensitivity level of data. In catalogs, a classification describes the sensitivity of a whole data asset to help catalog members understand the asset. You can use classifications to describe Business Terms, Data Classes, Reference Data Sets, and Governance Rules. In data protection Rules, you can include Classifications in conditions to identify the type of data to restrict.

__38. Click `Navigation Menu` ⇨ `Organize` ⇨ `Data and AI governance` ⇨ `Classifications.`

__39. Scroll to find Classification `Confidential` and click on it



__40. The Classification is described here. You could also add the primary Category here, but there is no need to do so now.



4.7.2 Data classes

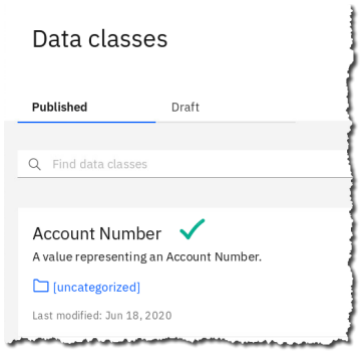
Data classes describe the contents of data in a column in a relational or structured data set. Data classes are assigned to columns when profiling a structured data asset and shown on the Profile page in a Catalog or Project.

Watson Knowledge Catalog provides a predefined set of Data Classes. Some Data Classes are categorized into groups, for example:

- If you select **Date**, it also includes **Date of Birth**.
- If you select **Driver's License**, it also includes all driver licenses listed below.

__41. Click `Navigation Menu` ⇨ `Organize` ⇨ `Data and AI governance` ⇨ `Data classes.`

__42. Scroll to find `Account Number` and click on it.



__43. The Data Class is described here.

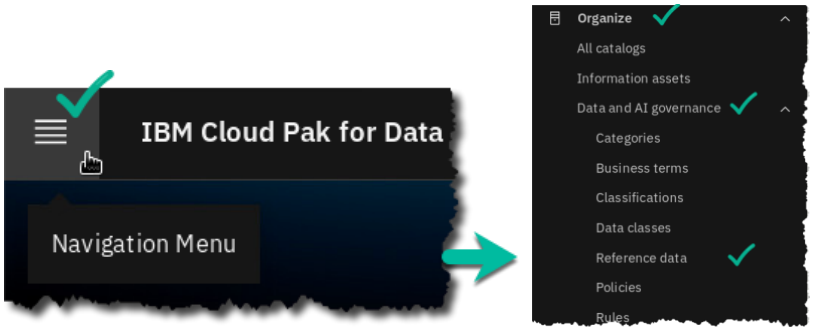
From here you could edit the description, examples or add the primary Category.

4.7.3 Reference Data Set

`Reference Data Sets` define list of permissible values that are allowed for use within a data field and may be referenced by Business Terms, Policies, Rules and Data Classes in Watson Knowledge Catalog.

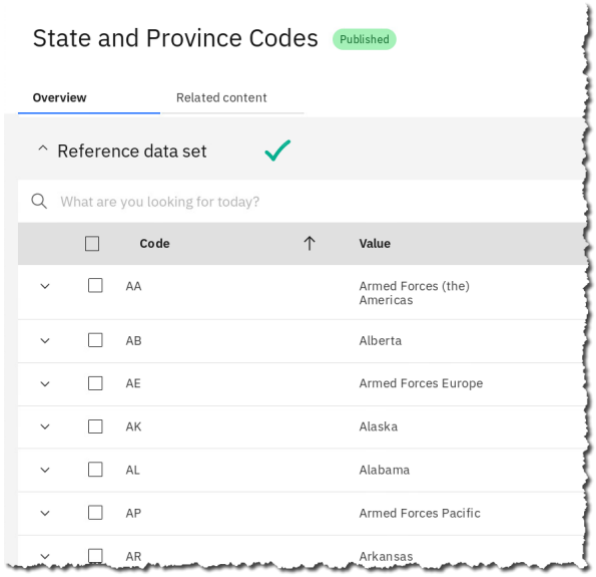
You can capture, manage, and socialize reference data — setting it up once and re-using the reference data in other places.

__44. Click `Navigation Menu` ⇨ `Organize` ⇨ `Data and AI governance` ⇨ `Reference data.`



__45. Scroll to find Reference Data Set `State and Province Codes` and click on it.

__46. Scroll down to review the Reference Data Set rows (the data).



__47. Click `Related content` and review how you could relate Data classes and Rules to it.

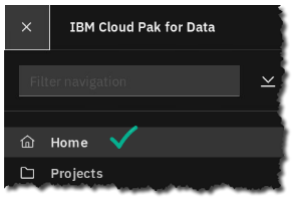
4.8 Searching for Data

The Data Scientist and Business Analyst personas may not always know what has been made available to them by the Data Engineers and Data Stewards in CPD. In fact, individual Data Engineers and Data Stewards may not always know what other users of the same persona have made available through their Collect and Organize activities.

This is remedied by the search functionality provided in CPD.

__48. Start the search here: `Navigation Menu` ⇨ `Home.`

(Doing this gives your search a neutral context.)

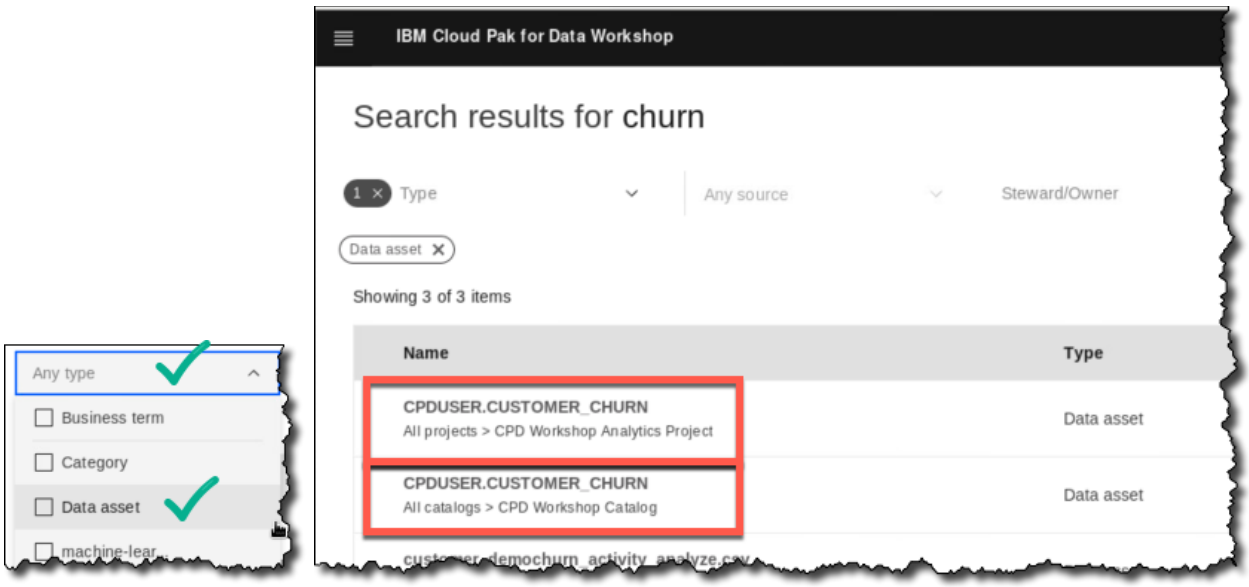


__49. Click in the `Search` bar.

__50. Type `churn` and hit `Enter`.

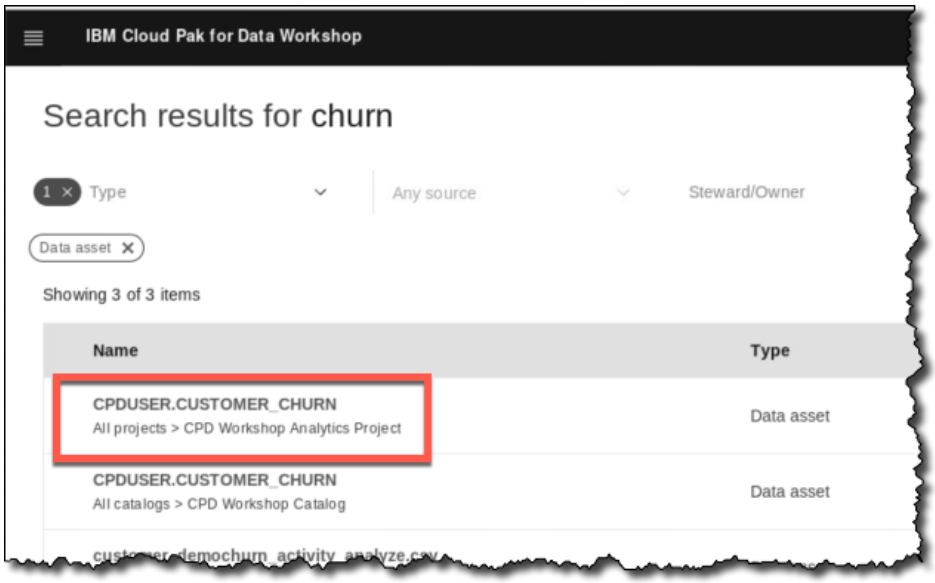
__51. Notice in the absence of a filter (or context) that multiple kinds of assets related to the term `churn` are found.

__52. Select from `Any type – Data Asset.`



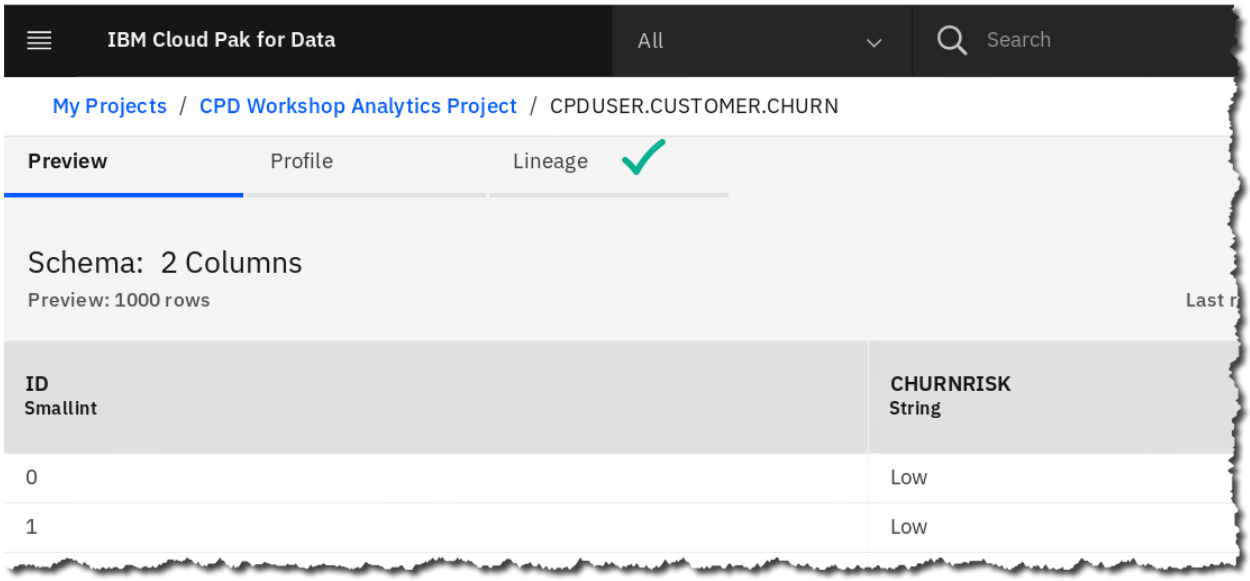
Notice that one Data asset is in a catalog and another is in a project.

__53. Click on the Data asset called `CPDUSER.CUSTOMER.CHURN` which is located in `All projects > CPD Workshop Analytics Project.`

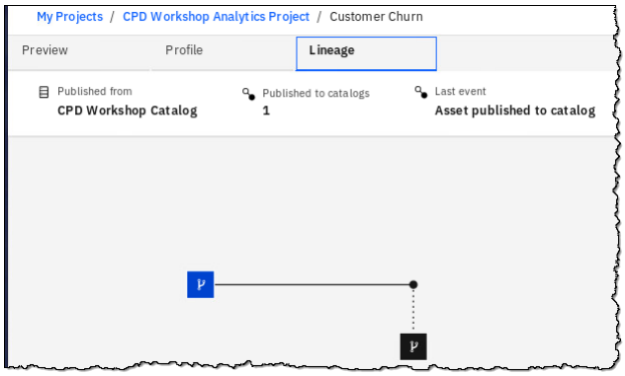


__54. Here you can review a sample of the data.

Next, click `Lineage.`



__55. The Lineage shown here is relatively simple for this Data asset; however, this could show a more complex lineage should this Data asset have been the result of a Data Flow Design or refine job flow output from a virtualized view of a join of two tables in two different servers.



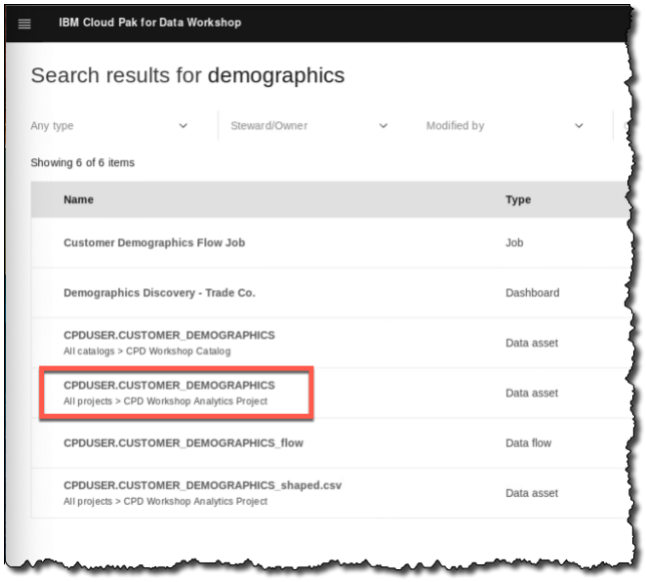
Data Steward

The results of this search and exploration of the Data asset tells us that we want to utilize this asset for a Data Flow Design job later in this lab.

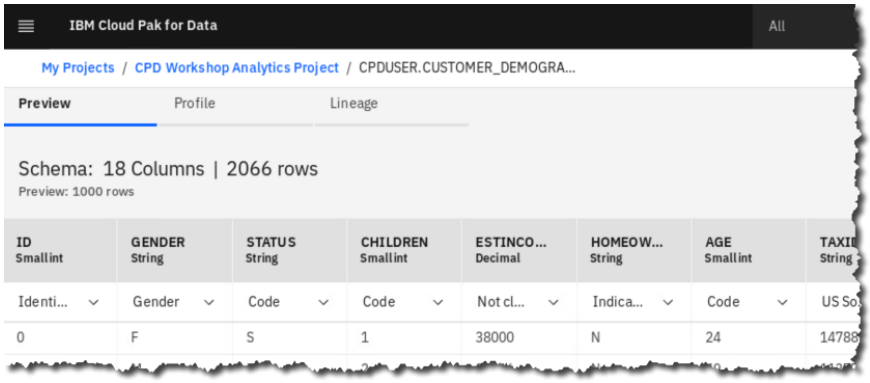
At this point, we could go to the `Overview` page of the `Catalog View` and click `Add to Project` for easy access later. However, this has already been done, so there is no need to do so now.

__56. Try the `Search` again, this time using the word `demographics`.

__57. Choose the second Data asset named `CPDUSER.Customer_Demographics` which is under `All projects > CPD Workshop Analytics Project.`



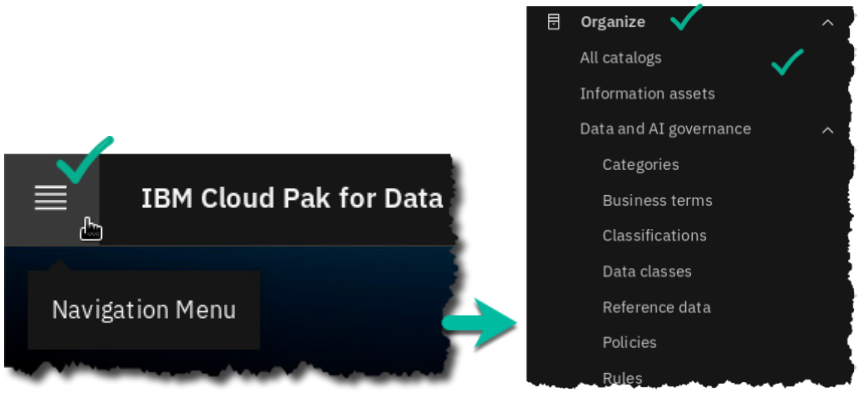
__58. This takes us to the `CPDUSER.Customer_Demographics` Data asset we reviewed in an earlier exercise at the beginning of this lab.



 Data Steward	You will be using this data in the Data Flow Design exercise to follow. To recap: you have searched for, found and selected two tables, from which you will build a job to join and transform this data to create another permanent table.

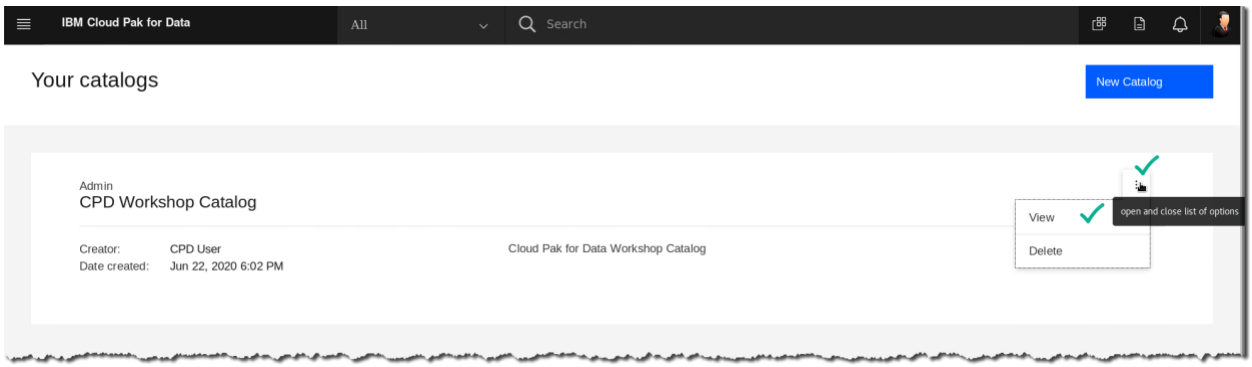
4.9 Reviewing the catalog

__59. Click `Navigation Menu` ⇨ `Organize` ⇨ `All catalogs.`

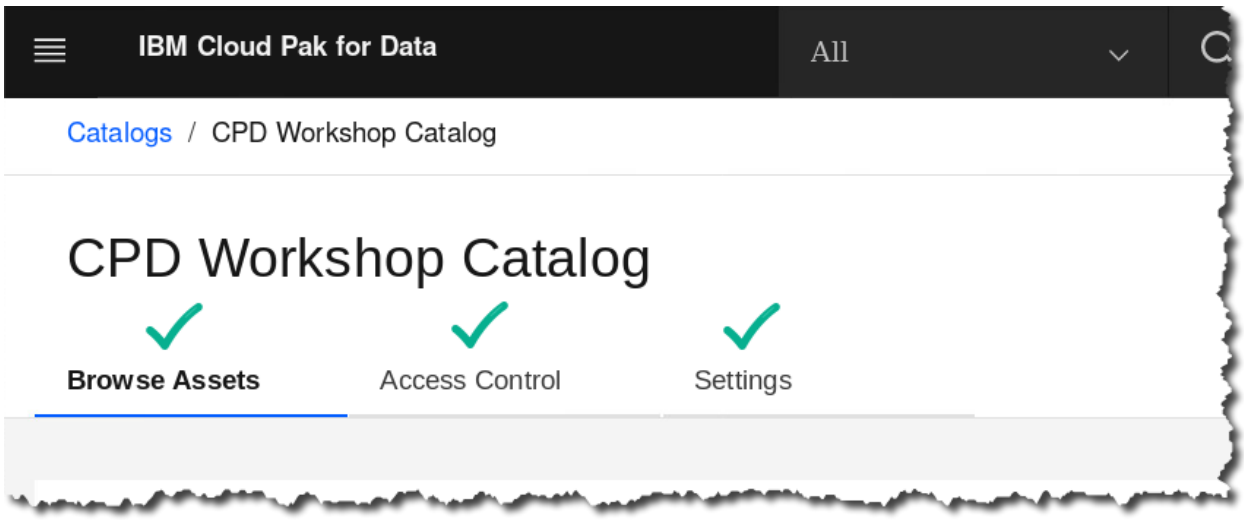


__60. You will see a catalog named `CPD Workshop Catalog`.

Review it by clicking `ellipses` ⋮ ⇨ `View.` *(Tip: You can also click on the catalog name.)*

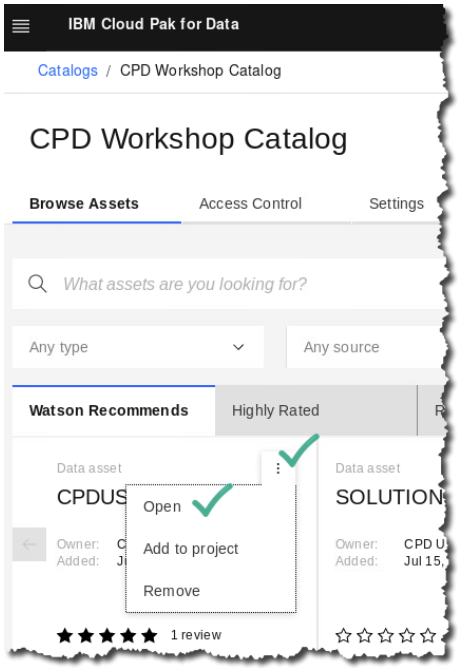


__61. Review each of the Catalog sections: `Browse Assets`, `Access Control`, and `Settings.`



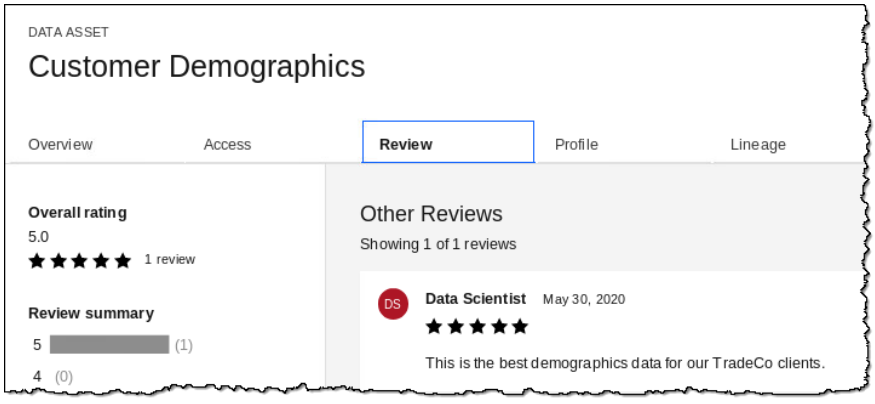
__62. Return to `Browse Assets.`

__63. Click the. ellipses to the right of the `CPDUSER.Customer_Demographics` tile.
(Tip: You can either hover over the name for a flyout that shows the complete name or zoom out.)



__64. Click the tab `Review.`

__65. This is where you can find review ratings and comments for the data, which is another aid in helping you find the best data for your projects.

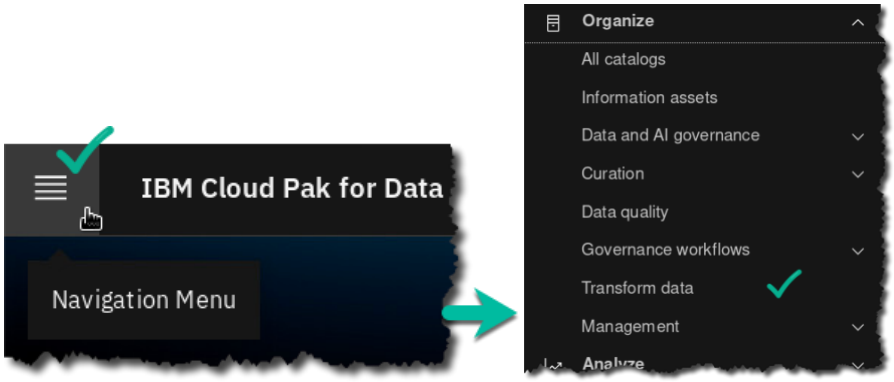


4.10 Transforming Data

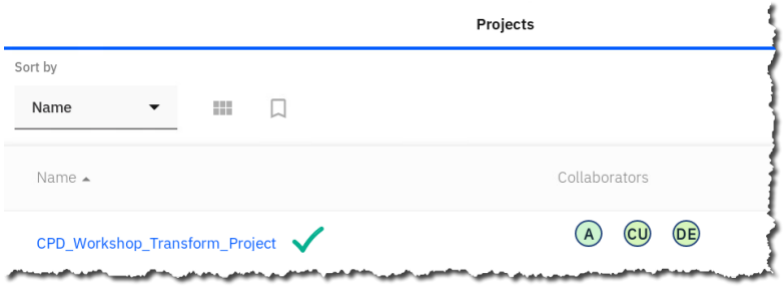
In this set of exercises, you will use the CPD built-in Data Flow Designer (DFD) to build a job that can transform your data. Note: it is the Data Engineer who would do this.

Persona (Role)	Capabilities
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

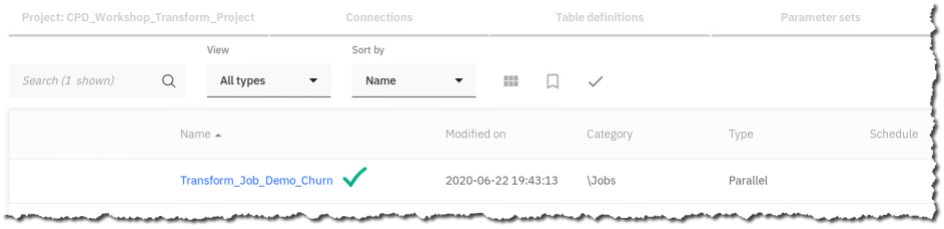
__66. Click `Navigation Menu` ⇨ `Organize` ⇨ `Transform data.`



__67. Select the project: `CPD_Workshop_Transform_Project.`

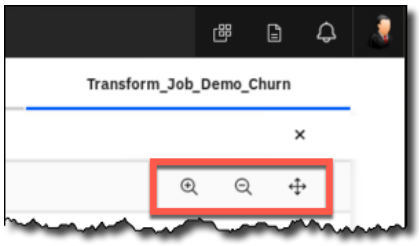


__68. Select the job `Transform_Job_Demo_Churn.`



 Data Engineer	Note: if there is a lock icon next to this job then click on the lock icon to unlock the job, then you can select it. Just make sure you are not in two different web client sessions at the same time doing this because that might be causing your locked job issue.

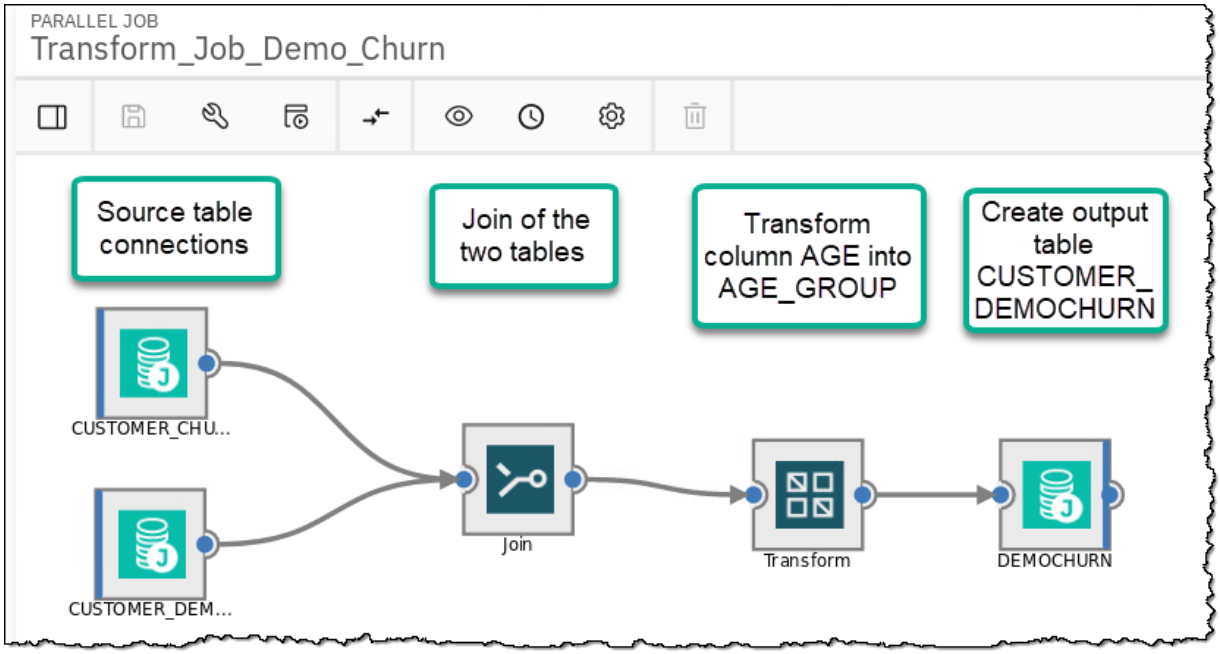
__69. Use the zoom icons to get the best view of the entire transform job on your screen.



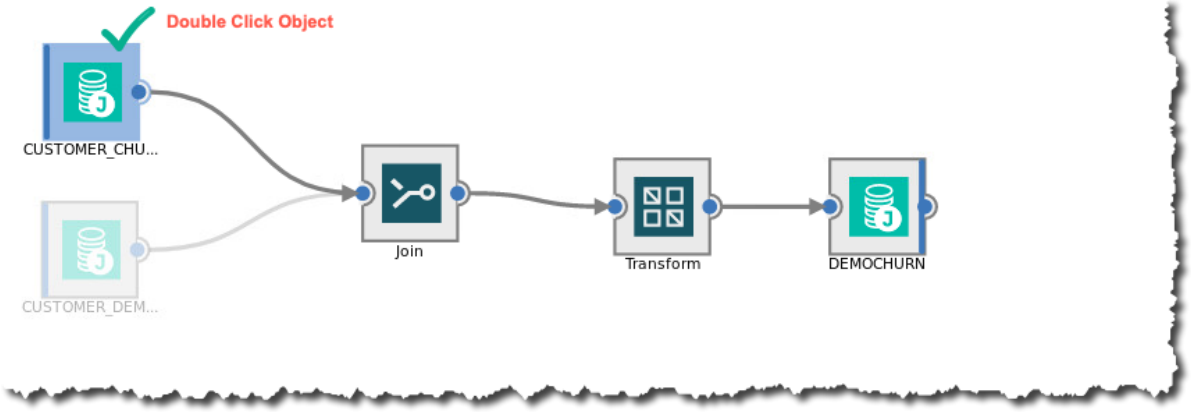
__70. This job consists of four stages as shown below. Its purpose is to join two tables (represented by the Data assets you previously searched for) `CUSTOMER_CHURN` and `CUSTOMER_DEMOGRAPHICS`.

It then adds one new column called `AGE_GROUP` which is derived (transformed) from the column `AGE`.

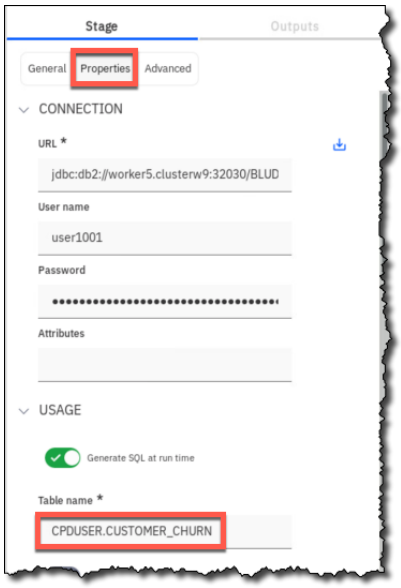
Finally, it writes out the results to a single table `CUSTOMER_DEMOCHURN`.



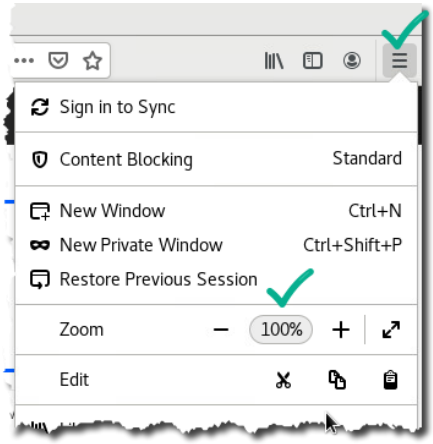
__71. Review the connection of CUSTOMER_CHURN (the top one) by double clicking on it.



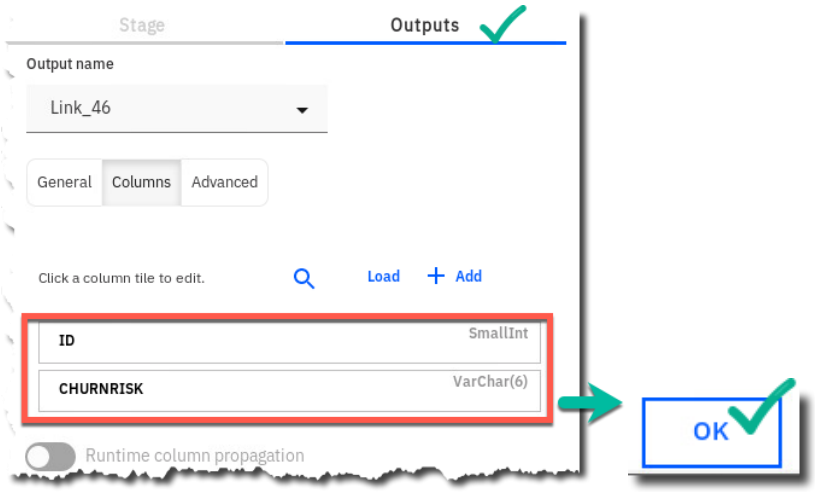
__72. The `Properties` for this connection will appear. Review the connection and other properties.



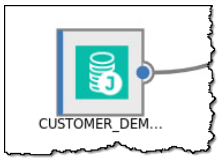
(Note: if this does not appear, resize your browser zoom as shown below)



__73. Review the tab `Outputs` to see the table columns being passed on to the next stage of the job. Then click `OK`.

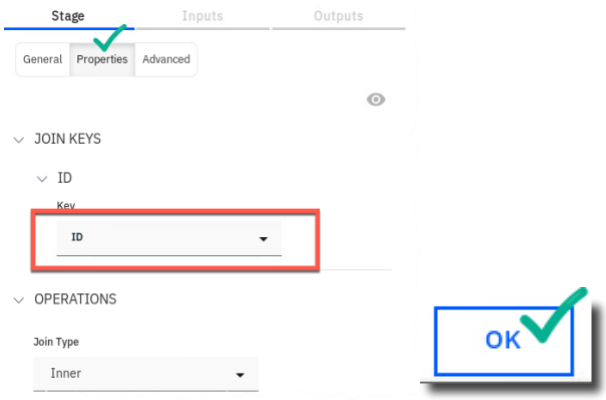


__74. Do the same with the connection for `CUSTOMER_DEMOGRAPHICS` (the bottom one).



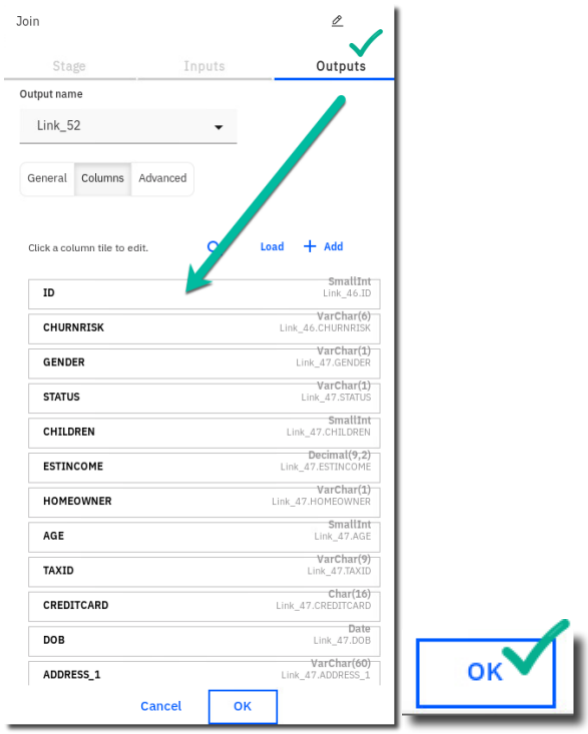
__75. Next, double click on the stage `Join.`

__76. Notice in this stage's `Properties`, the two tables are joined by the column: `ID.`

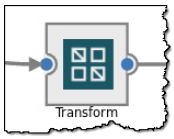


__77. In this Join stage, select section: `Outputs.`

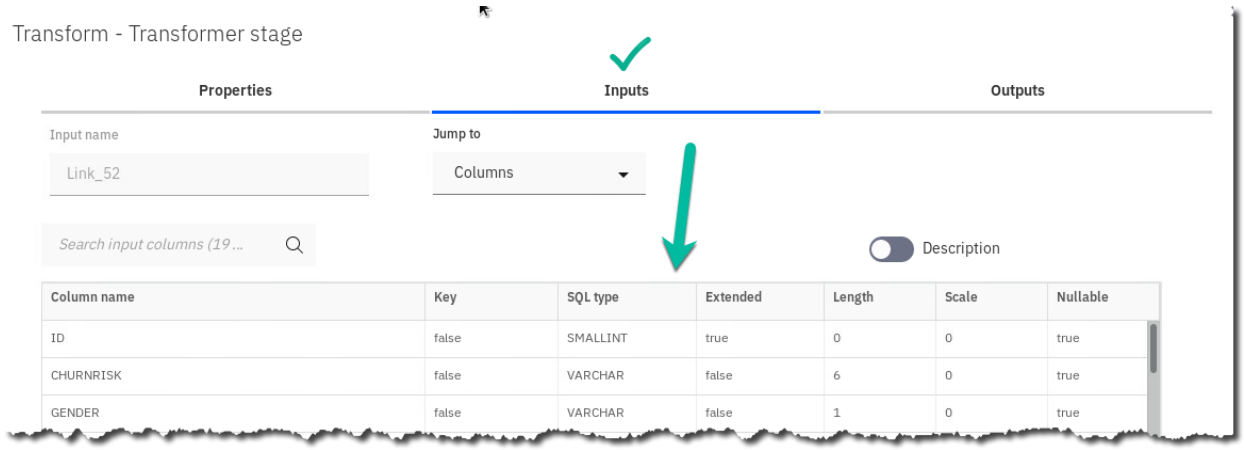
Notice that the columns now include both tables.



__78. Double click stage `Transform.`



__79. Select the section `Inputs`. These are simply the inputs from the Join stage.

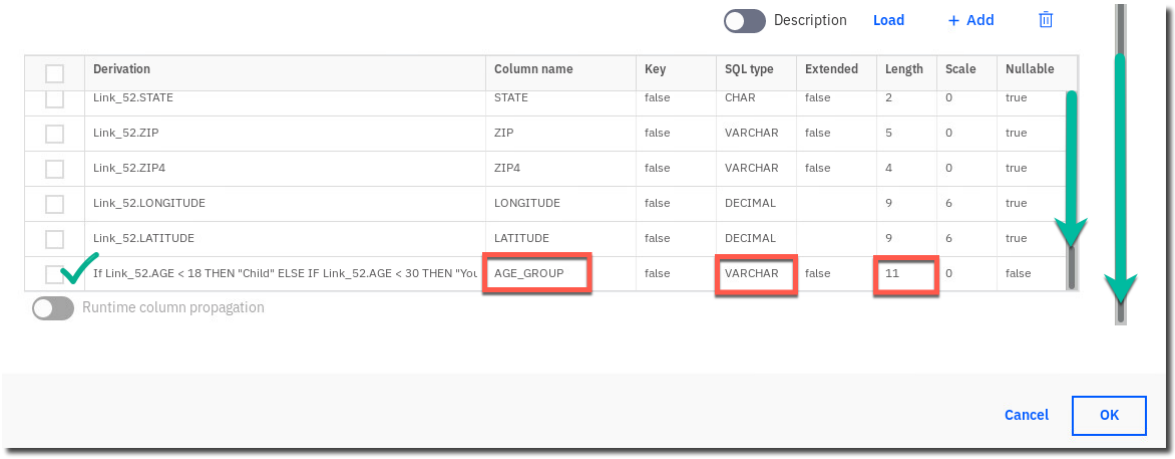


__80. Review the section `Outputs`. Notice each column from Inputs is given a coded “link” prefix before the column name.

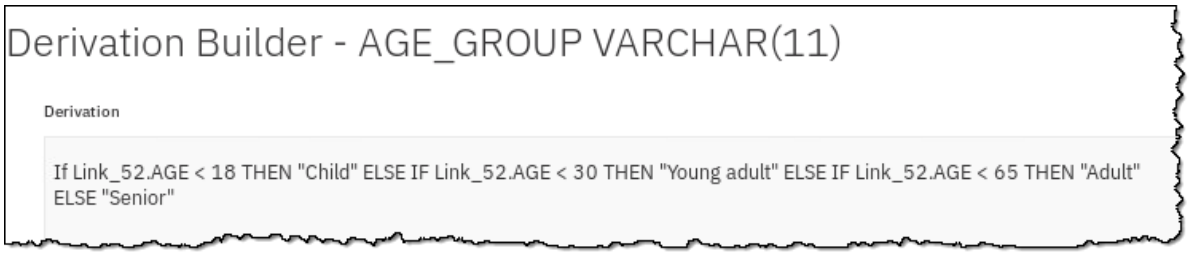
__81. Scroll down the list of output columns (using both scroll bars) to find the last output column, which is a derived (transformed) column called AGE_GROUP.

Notice it is assigned a SQL type (data type) or VARCHAR, Length 11.

Click on this derived column to review its formula.

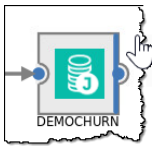


__82. The Derivation Builder screen shows how it transforms AGE into AGE_GROUP.



__83. Click Cancel and Cancel again to ensure you have not inadvertently changed something.

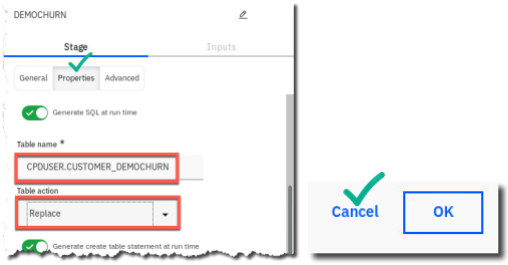
__84. Double click the last stage, DEMOCHURN



__85. Scroll down in the Properties to find `Table name` and `Table action.`

This indicates that `CPDUSER.CUSTOMER_DEMOCHURN` will be replaced with each run of this job.

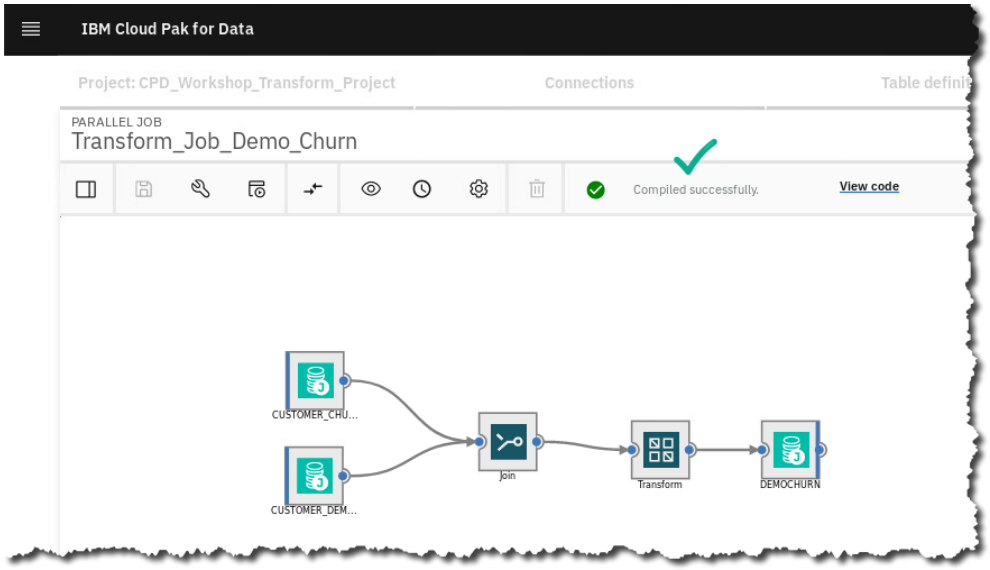
Close without changing anything in this stage by clicking `Cancel.`



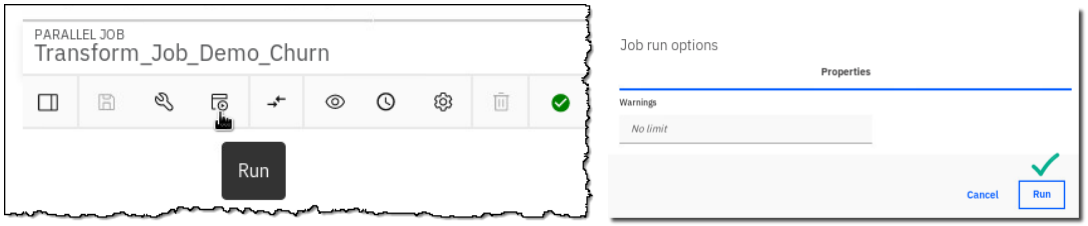
__86. Click on the `Save` then `Compile` (wrench) `icon`.

 Data Engineer	The compile action does the following: • *Primary Input.* If you have more than one input link to a Transformer stage, the compiler checks that one is defined as the primary input link. • *Reference Input.* If you have reference inputs defined in a Transformer stage, the compiler checks that these are not from sequential files. • *Key Expressions.* If you have key fields specified in your column definitions, the compiler checks that there are key expressions joining the data tables. • *Transforms.* If you have specified a transform, the compiler checks that this is a suitable transform for the data type.

__87. The job should complete successfully.

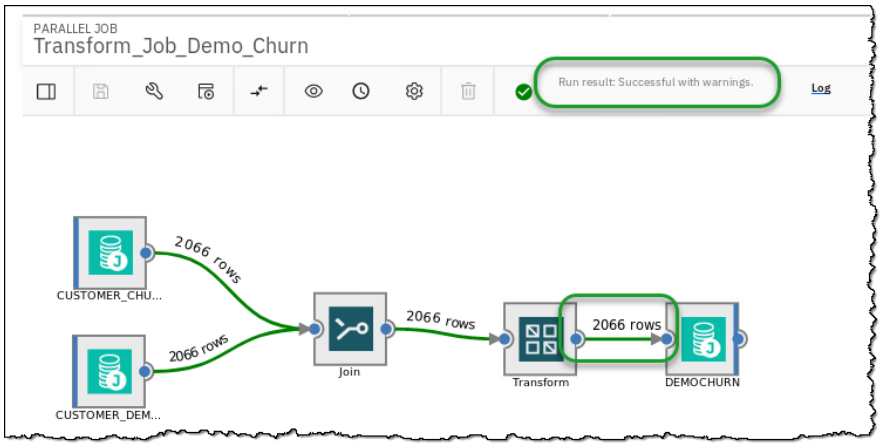


__88. Click the `Run icon`. When the Job run options box appears, select `Run.`



__89. Watch the job run…

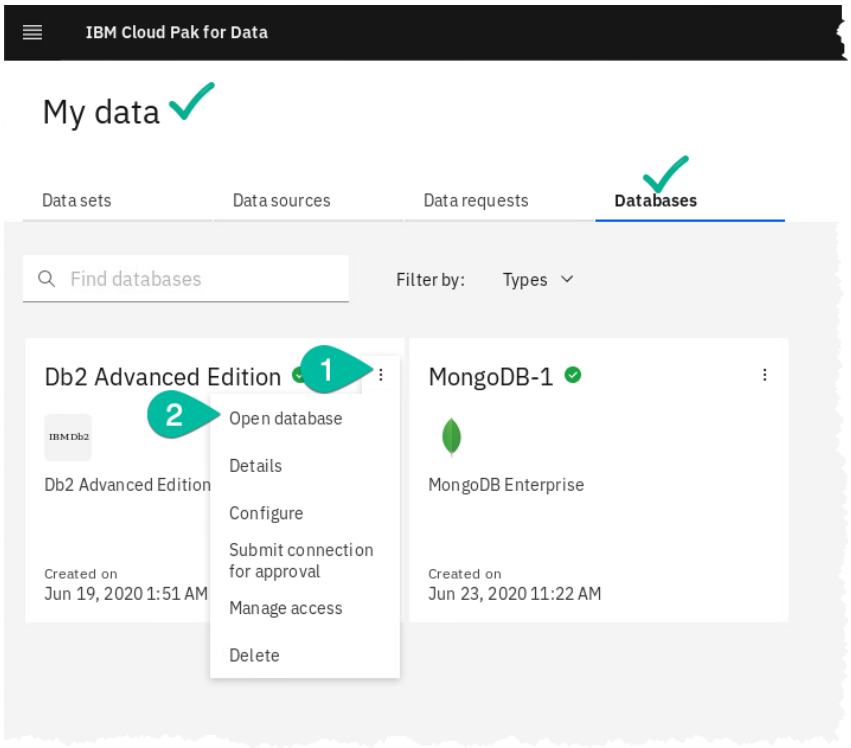
When it completes, it should look like this:



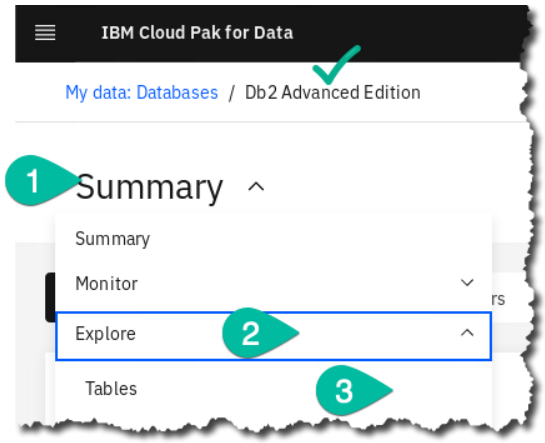
 Data Engineer	• You can select the Log to view any warnings or errors during this process. This is helpful when troubleshooting. In this case, you notice a warning that the script attempted to drop a table that didn't exist. No need to worry.

__90. Review the data by opening the Data Server Manager (DSM) for the Db2 Warehouse.

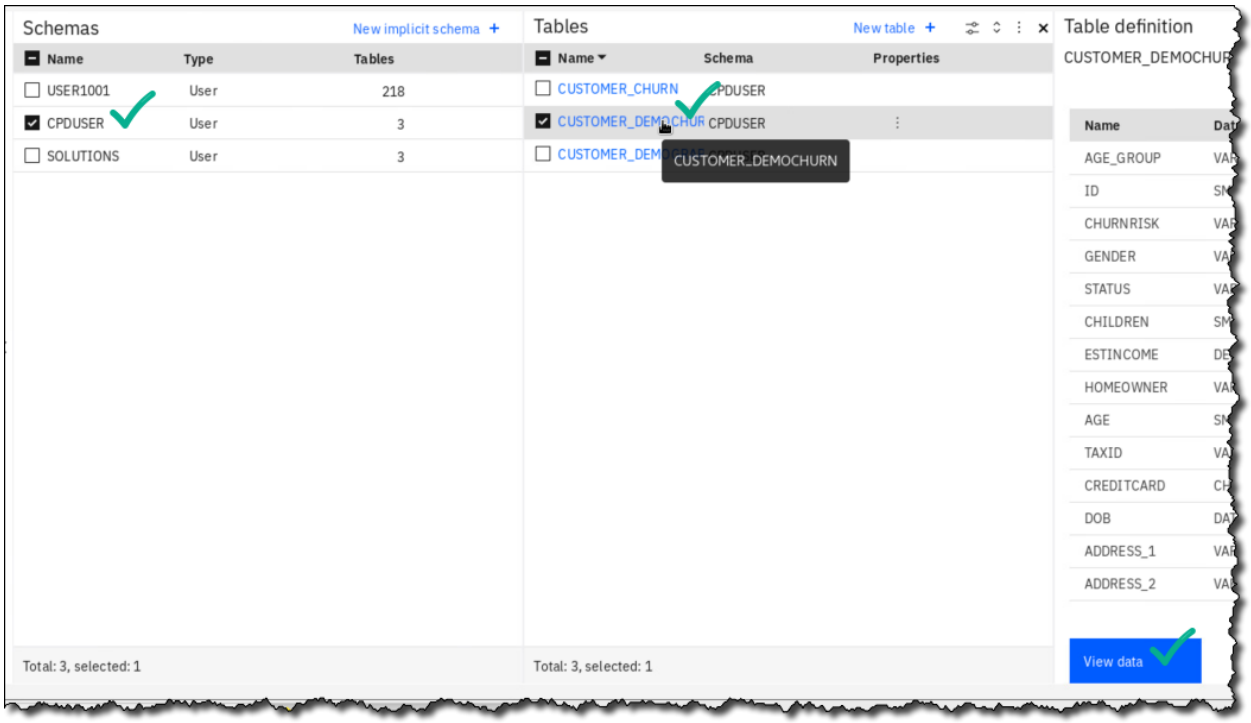
`Navigation Menu` ⇨ `Collect` ⇨ `My Data` ⇨ `Databases` ⇨ `Db2 Warehouse` ⇨ `ellipsis` ⇨ `Open database`



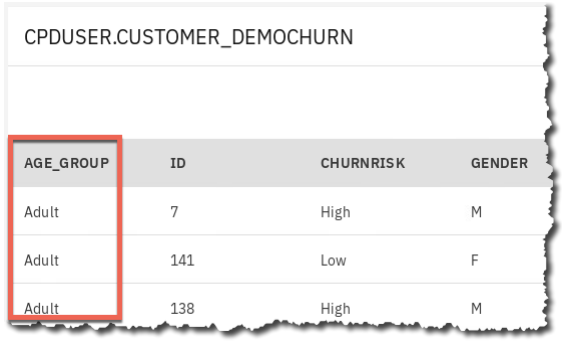
__91. At the top left of the Db2 DSM console, click on: `Summary` ⇨ `Explore` ⇨ `Tables.`



__92. Select schema `CPDUSER` ⇨ table `CUSTOMER_DEMOCHURN` ⇨ `View Data.`



__93. The new column `AGE_GROUP` contains the derived data, and the output is also the join of the two tables in this schema.



CPDUSER.CUSTOMER_DEMOCHURN

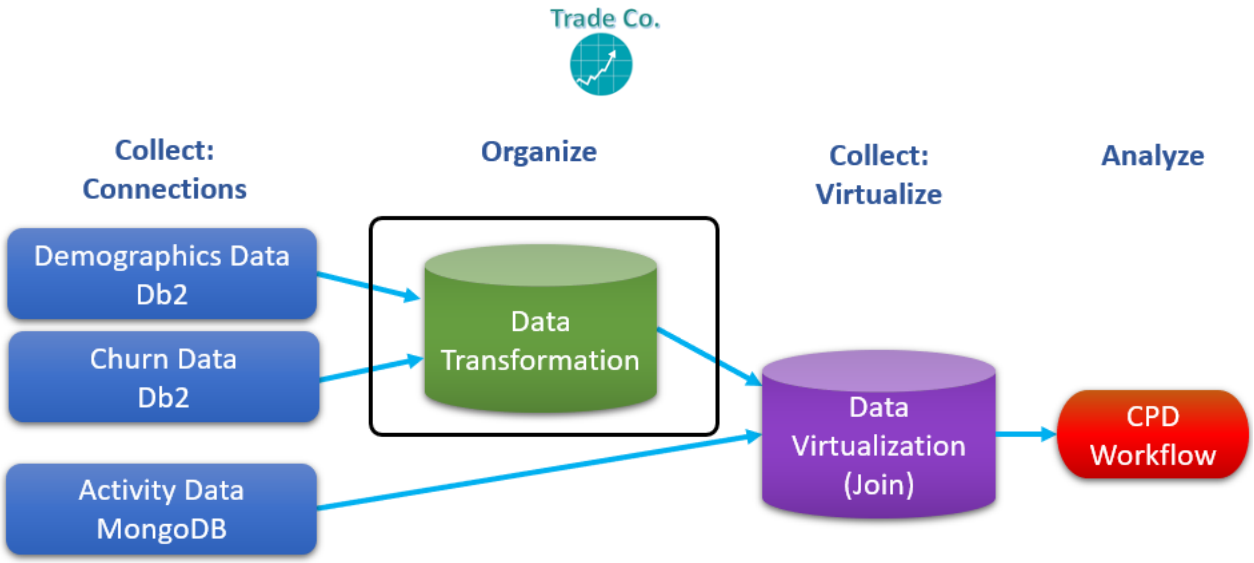
AGE_GROUP	ID	CHURNRISK	GENDER
Adult	7	High	M
Adult	141	Low	F
Adult	138	High	M

 Data Engineer	If you want to try your hand at building a transformation job from scratch, please see the Organize Deeper Dive lab, section “Transformation Data – Creating a job.”

4.11 Lab conclusion

We have seen the value in creating a Data Dictionary by creating a glossary of categories and terms, policies and rules, to make data searchable so that data scientist, data engineers and business analysts can “shop for data.”

This lab showed you how data can be profiled, visualized, refined, searched and transformed.



The steps covered here could normally take many weeks, months, and sometimes even years, to complete using traditional manual methods. Cloud Pak for Data automates these things so that you can accelerate the time to value of your organization’s analytics projects.

**** End of Lab 04 - Organize**

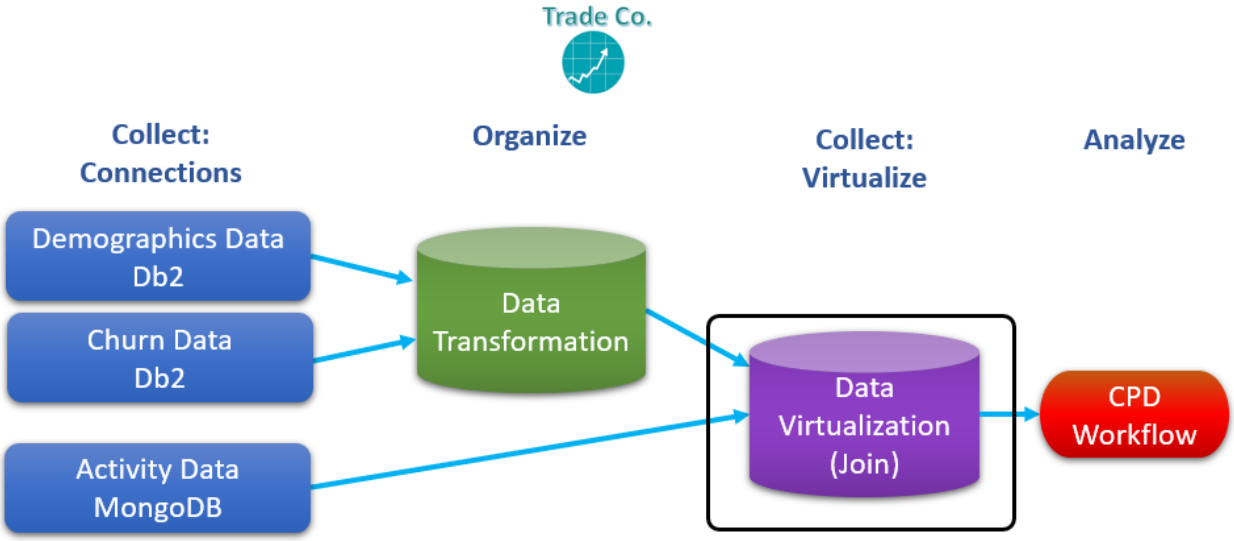
Lab by Burt Vialpando, Kent Rubin and John Van Buren, IBM

Lab 05 COLLECT: VIRTUALIZE

5.1 Lab overview

In this lab, you will learn about `Data Virtualization` to complete the `Collect` tasks by creating a virtualized view of the transformed Db2 Demographics and Churn data, joined with the MongoDB Activity data.

The team from Trade Co. wants to use data virtualization to easily join the disparate data sources into one view for easier consumption in their analytics work.



5.2 Persona represented in this lab

The `Data Engineer` persona will be performing the various `Collect` tasks shown in this lab.

Persona (Role)	Capabilities
 Data Engineer	Data Engineers build and optimize the systems to allow data scientists and business analysts to perform their work. The Data Engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

5.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

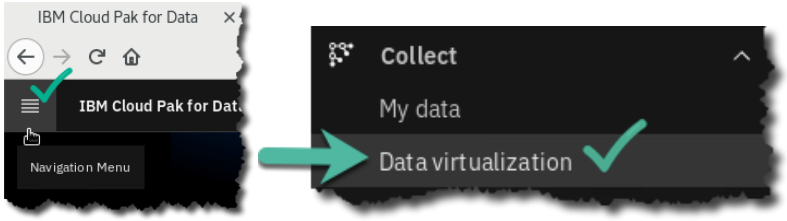
__2. Double-click the desktop icon: `Cloud Pak for Data Web Client.`

__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

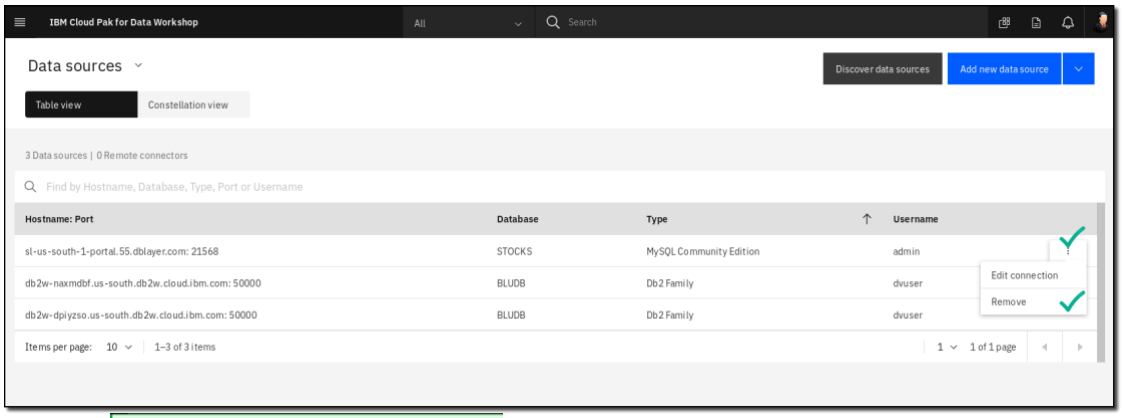
5.4 Remove older Data Virtualization sources

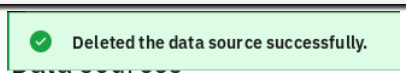
This step is needed to remove older data sources that are no longer required for this lab.

__4. Start at the `Navigation Menu` (“hamburger” icon) ⇨ `Collect` ⇨ `Data virtualization.`



__5. In the Data sources screen, click the `ellipse` on the MySQL Community Edition and `Remove`

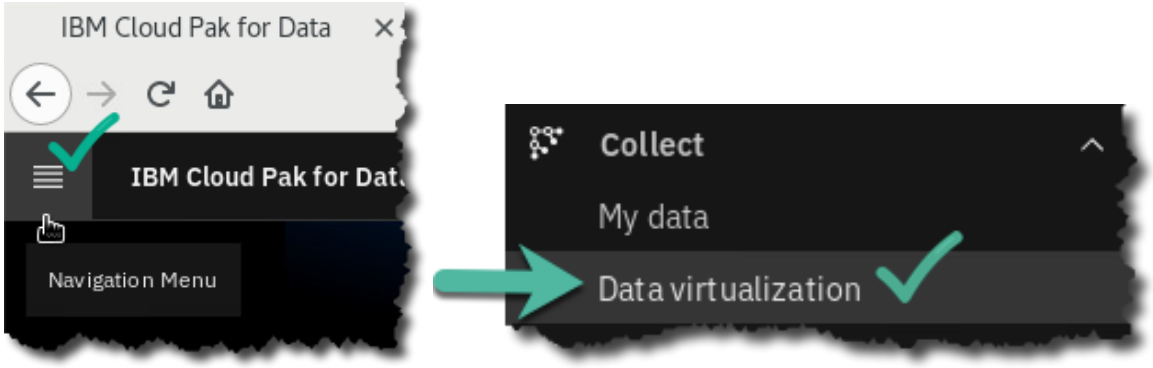




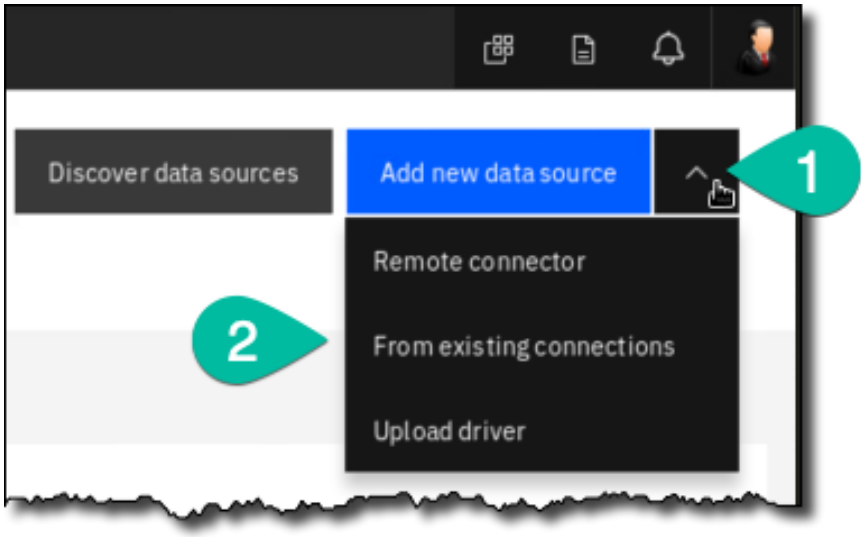
5.5 Adding Data Virtualization data sources

Let's explore the Data Virtualization process by adding new data sources to virtualize.

__6. Start at the `Navigation Menu` ("hamburger" icon) ⇨ `Collect` ⇨ `Data virtualization.`



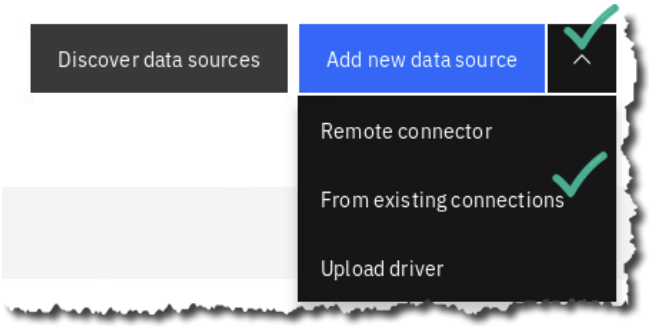
__7. In the Data sources screen, click the `Arrow` next to 'Add new data source' ⇨ `From existing connections.`



__8. Select (click on) `Db2 Advanced Edition` ⇨ `Next.`

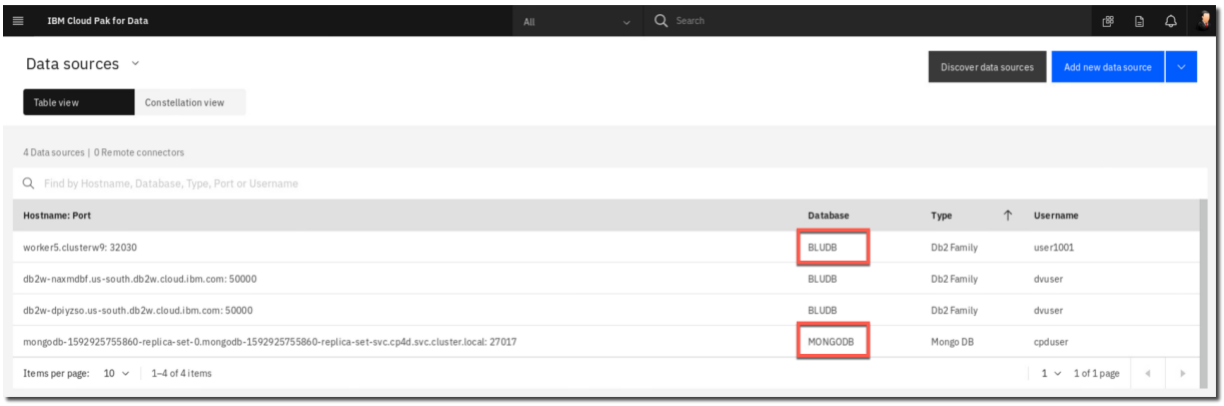
__9. You will notice a new Data source hostname created by Username user1001 (port 32030)

__10. Repeat this again: `Add new data source` ⇨ `From existing connections.`



__11. Select (click on) `MongoDB` ⇨ `Next.`

__12. You should now see the following Data virtualization Data sources.

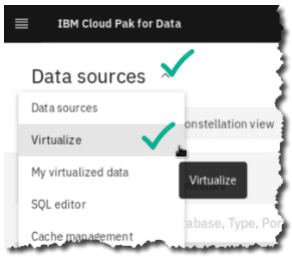


Data Engineer	In this exercise, you used previously created connections to create Data sources in the Data virtualization instance. You could have created the connection dynamically on this screen as well.

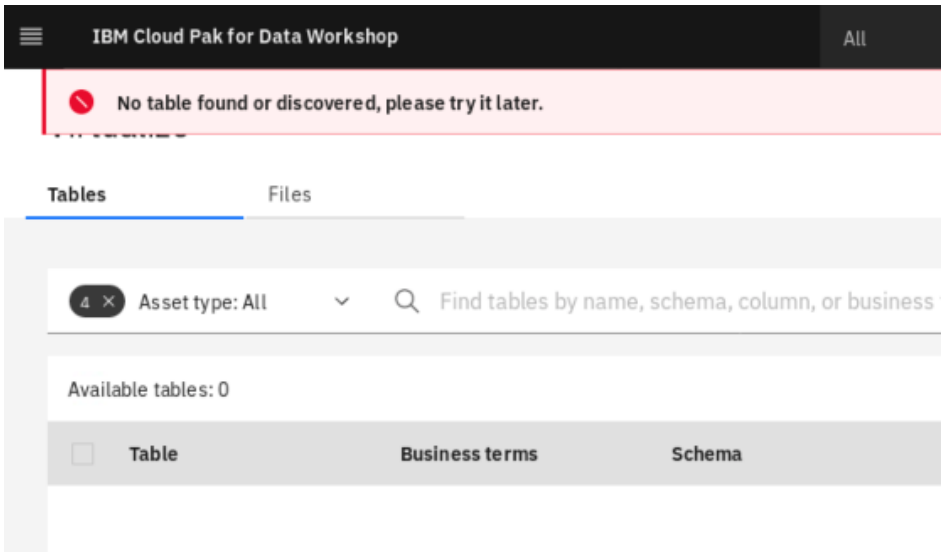
5.6 Virtualizing the Db2 data

In the next exercise, you will virtualize the Db2 table `CUSTOMER_DEMOCHURN`.

__13. At the top left of the screen, click `Data sources` ⇨ `Virtualize.`



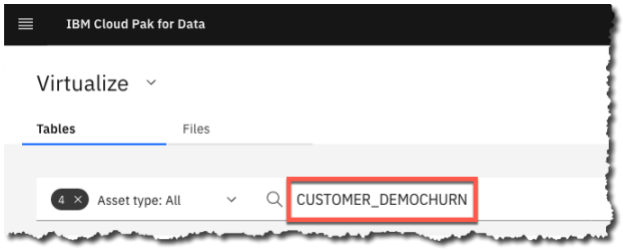
NOTE: If you experience an error message after a minute or two and see this message:



Simply wait 5 minutes then refresh the browser and the tables should appear.

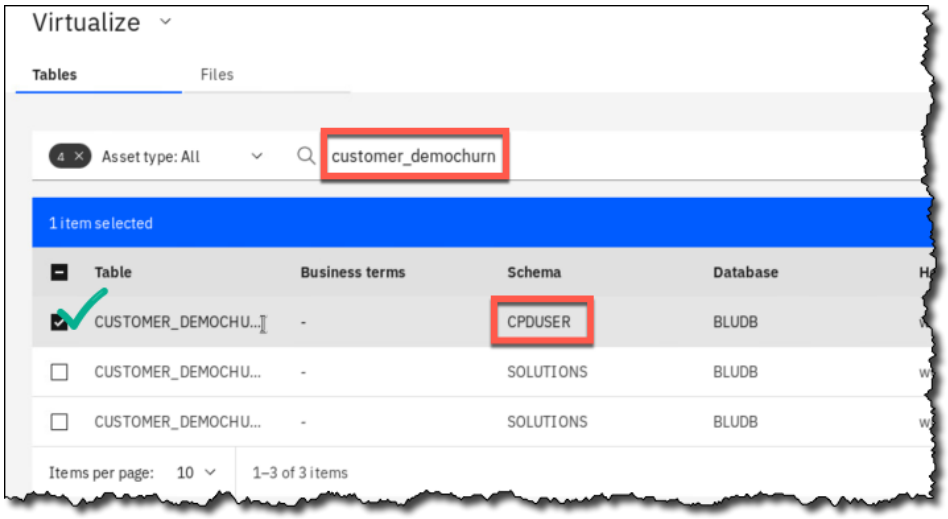
__14. Under `Databases`, check box `Db2 Family.`

Fill in the `Search` bar with `CUSTOMER_DEMOCHURN.`



__15. Check the able `CPDUSER.CUSTOMER_DEMOCHURN` from the `CPDUSER` schema.

If. You didn't do the Organize/Transform lab, choose `SOLUTIONS.CUSTOMER_DEMOCHURN.`



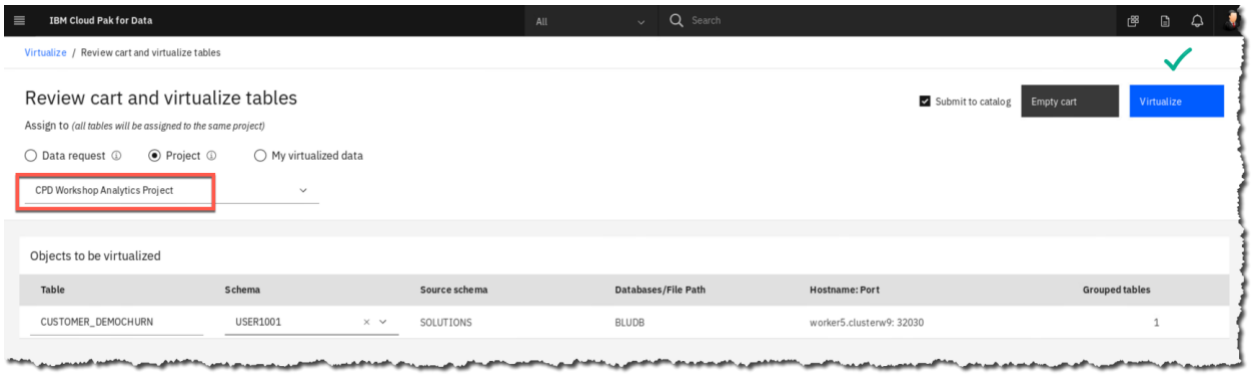
Note: the above tables are identical, but the first one was only created if you did the Organize / Transform lab preceding this lab. The pre-created table is useful for jumping right into this lab. Also, use your browser functionality to zoom out if you cannot see the full table names clearly.

__16. Click `Add to Cart.`

__17. Click `View cart.`

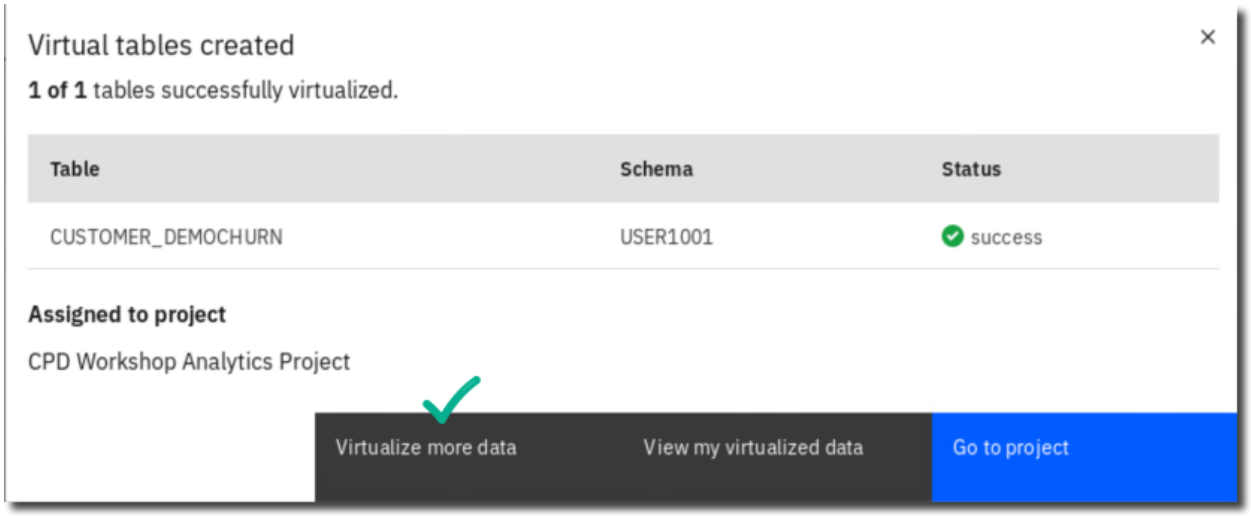
__18. Select *Project* `CPD Workshop Analytics Project.`

Click `Virtualize`



__19. You have now just created a virtualized table and placed it in a project.

Click `Virtualize more data.`

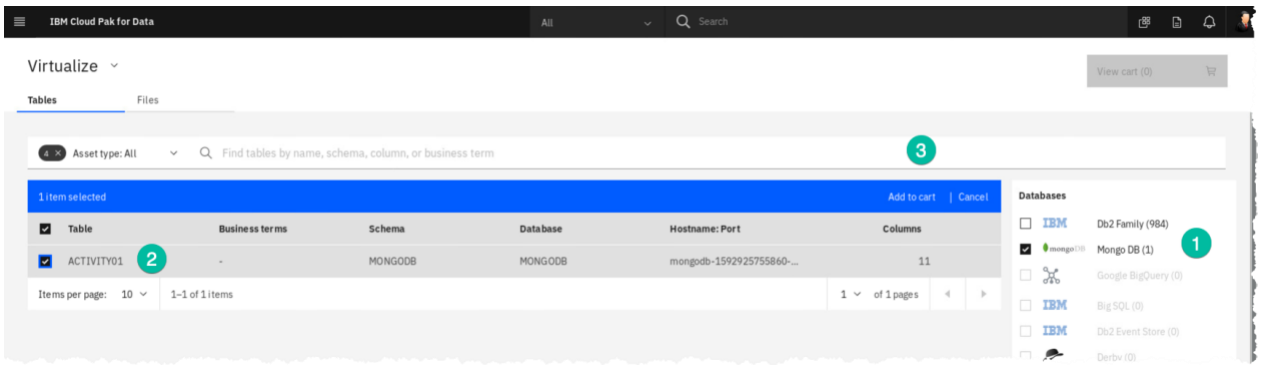

Virtual tables created
1 of 1 tables successfully virtualized.
Table
Schema
Status
CUSTOMER_DEMOCHURN
USER1001
success
Assigned to project
CPD Workshop Analytics Project
Virtualize more data
View my virtualized data
Go to project

5.7 Virtualizing the MongoDB data

You should be at the `Virtualize` screen again, after clicking `Virtualize more data` from the previous step.

If you are not there, you can always get there by doing the following: `Navigation menu` ⇨ `Collect` ⇨ `Data virtualization` ⇨ `Menu` ⇨ `Virtualize.`

__20. In the *Virtualize* screen Click the *Databases* box `Mongo DB`, select `ACTIVITY01.`



__21. Click `Add to cart.`

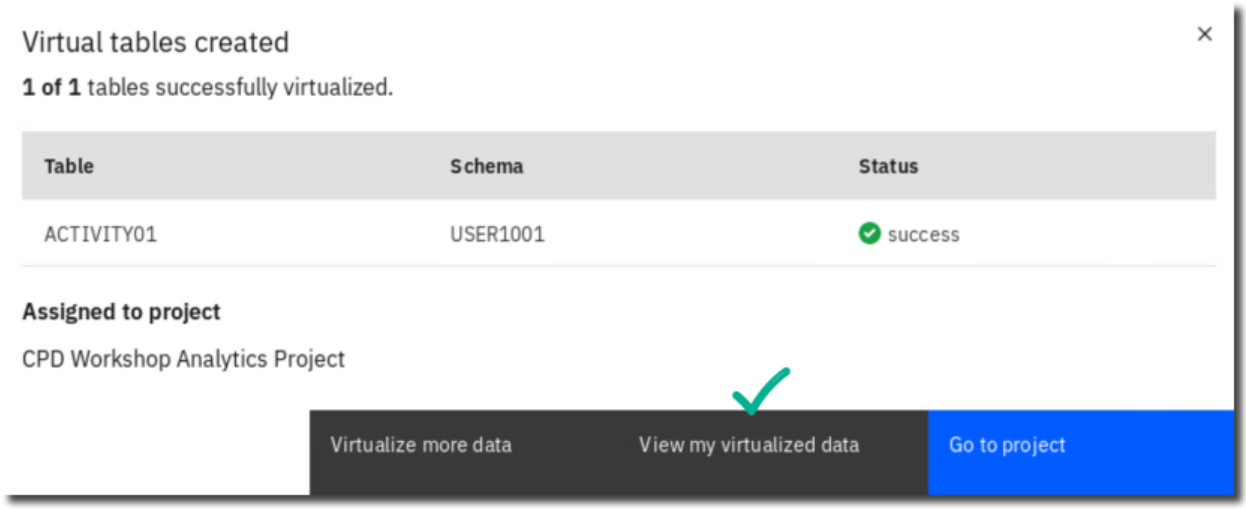
__22. Click `View cart.`

__23. Select *Project* `CPD Workshop Analytics Project.`

Click `Virtualize.`

__24. You have now created a virtualized table of Mongo data in a project.

Click `View my virtualized data.`



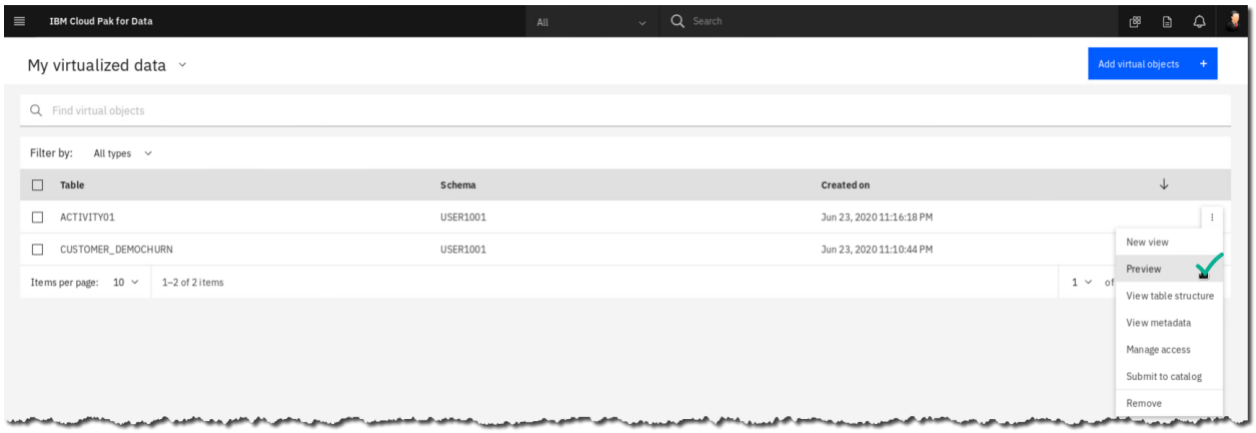
5.8 Joining the virtualized tables

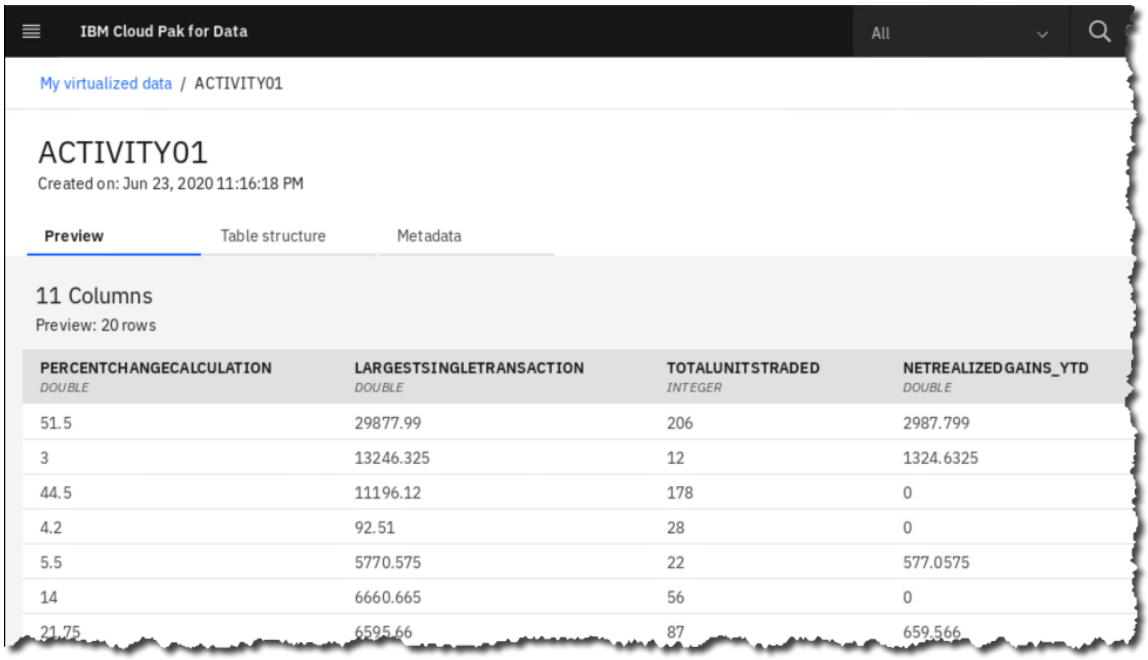
You should be at the `My Virtualize data` screen, after clicking `View my virtualized data` in the previous exercise.

If you are not, you can always navigate there by doing the following: `Navigation menu` ⇨ `Collect` ⇨ `Data virtualization` ⇨ `Menu` ⇨ `My Virtualized data.`

__25. Before joining any of the virtualized tables you have created, preview each of them to make sure they are working properly.

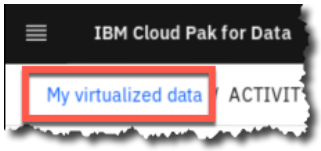
For each table, click on `ellipses` ⋮ ⇨ `Preview.`



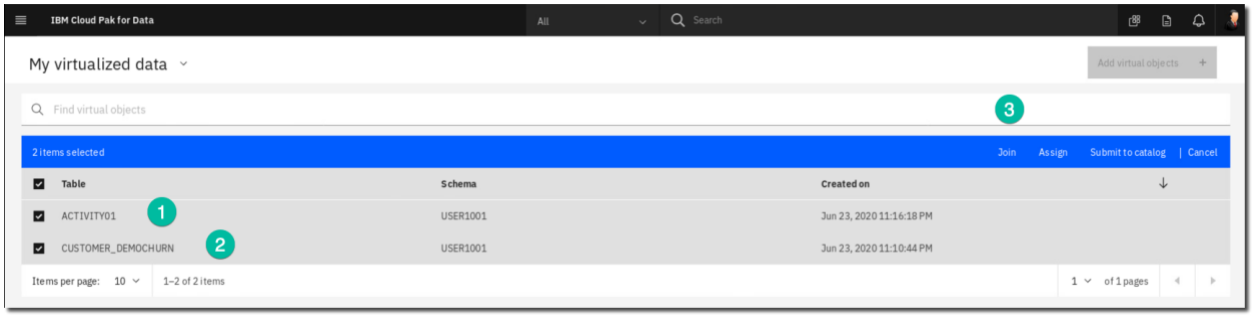


 Data Engineer	Check for data in ACTIVITY01 and CUSTOMER_DEMOCHURN. If any of these virtualized tables are not displaying data, you can fix it by performing a Remove action on the virtualized table (use the ellipses ⋮) and then re-create it.

__26. Return to My virtualized data by selecting breadcrumb back.



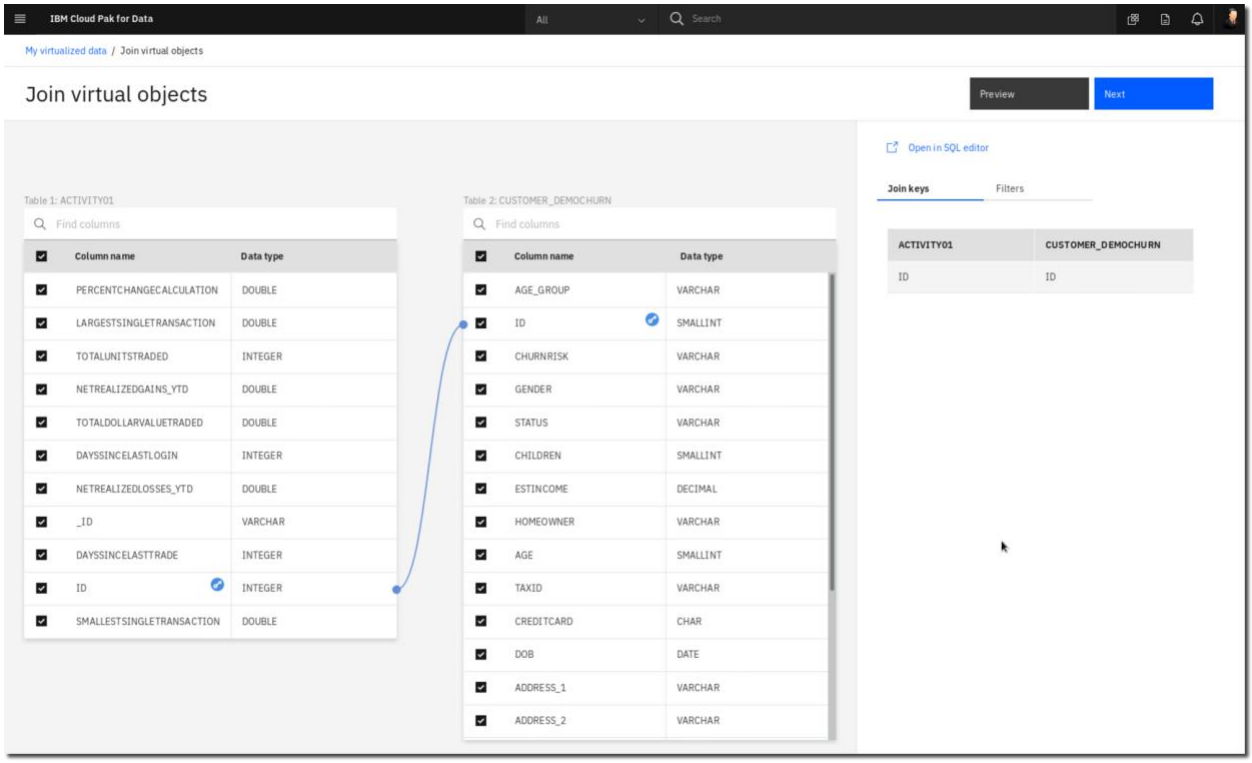
__27. Choose tables `ACTIVITY01` and `CUSTOMER_DEMOCHURN`, then click `Join.`



__28. Scroll down on the table `CUSTOMER_DEMOCHURN` until you can see the Column `ID.`

Join the virtualized tables by clicking on Column `ID` on the table `ACTIVITY01` and dragging the cursor to the Column `ID` on the table `CUSTOMER_DEMOCHURN`.

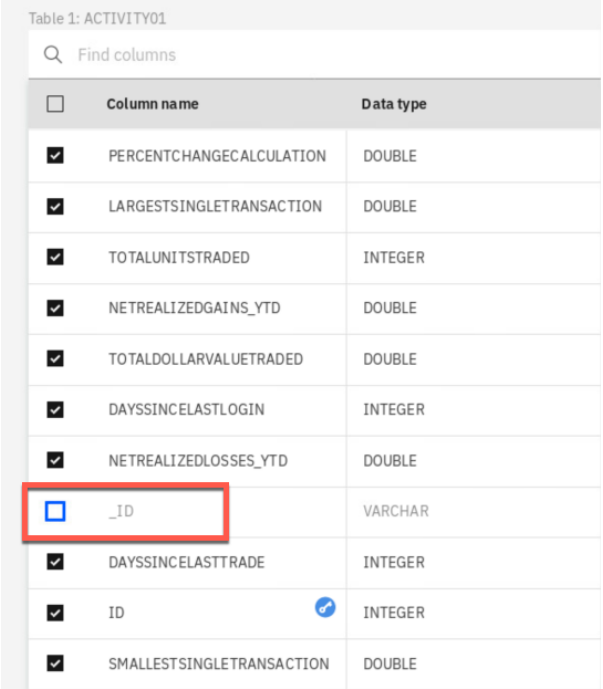
If done correctly, you will see blue key icons next to each column in the virtualized tables.



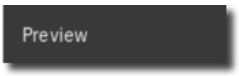
__29. On table `ACTIVITY01` find the Column `_ID`.

This column is different from the join column you just used called `ID.` The `_ID` field is a Mongo generated column that is not useful for our Analytics processing.

Click on the box next to the Column Name to uncheck Column `_ID.` This will indicate that you do not want this column included in the joined virtualized view of the two virtualized tables. You may have noticed that the arrow between ACTIVITY01 and CUSTOMER_DEMOCHURN disappeared. That's ok. The key still indicates a join relationship.

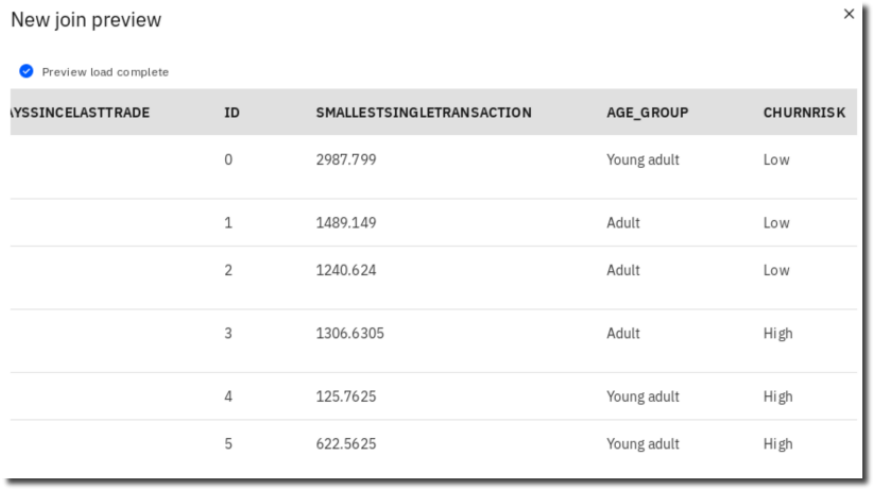


__30. Click `Preview.`



__31. Scroll to the right to make sure data columns from both tables are represented.

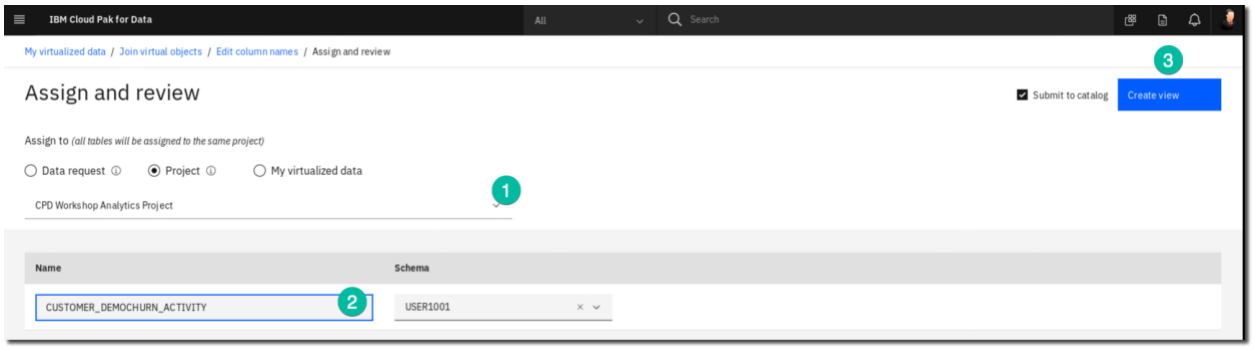
Click `x` to close.



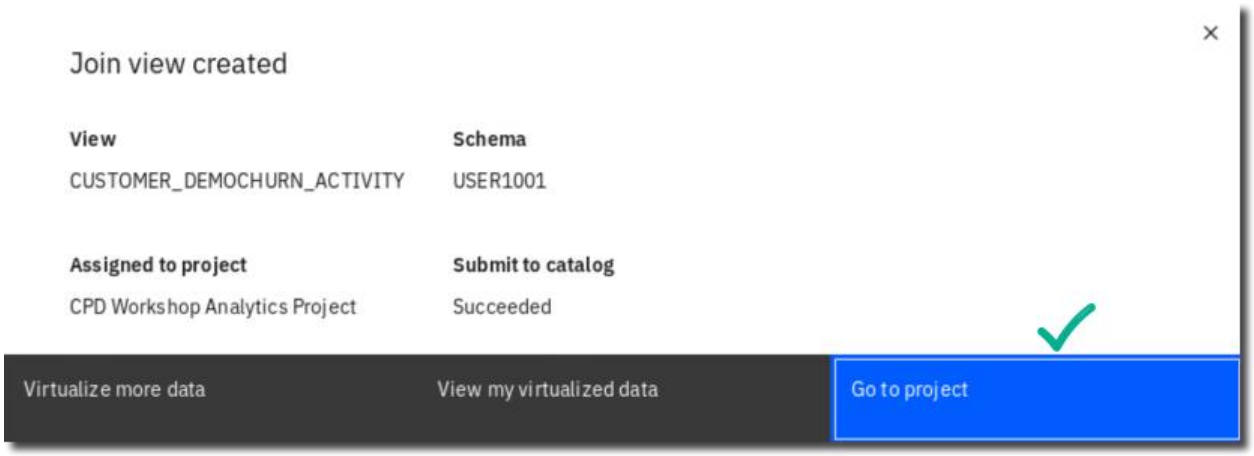
__32. Click `Next`. Now you can Edit column names if you like.

__33. Click `Next again.`

__34. In the Assign and review screen, select Project `CPD Workshop Analytics Project` (1), then type view name: `CUSTOMER_DEMOCHURN_ACTIVITY` (2) and Click `Create View` (3)

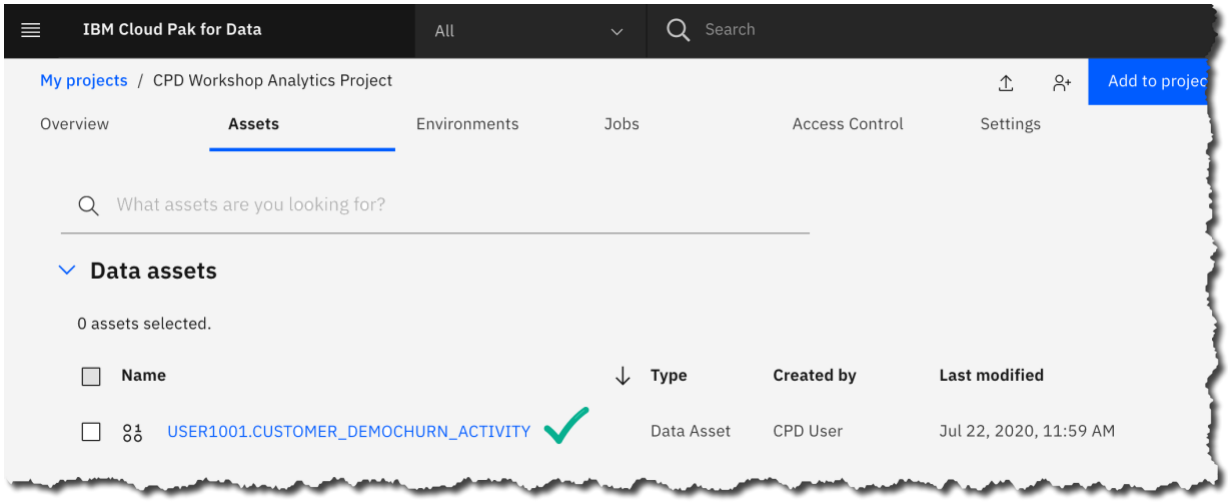


__35. Click `Go to project.`



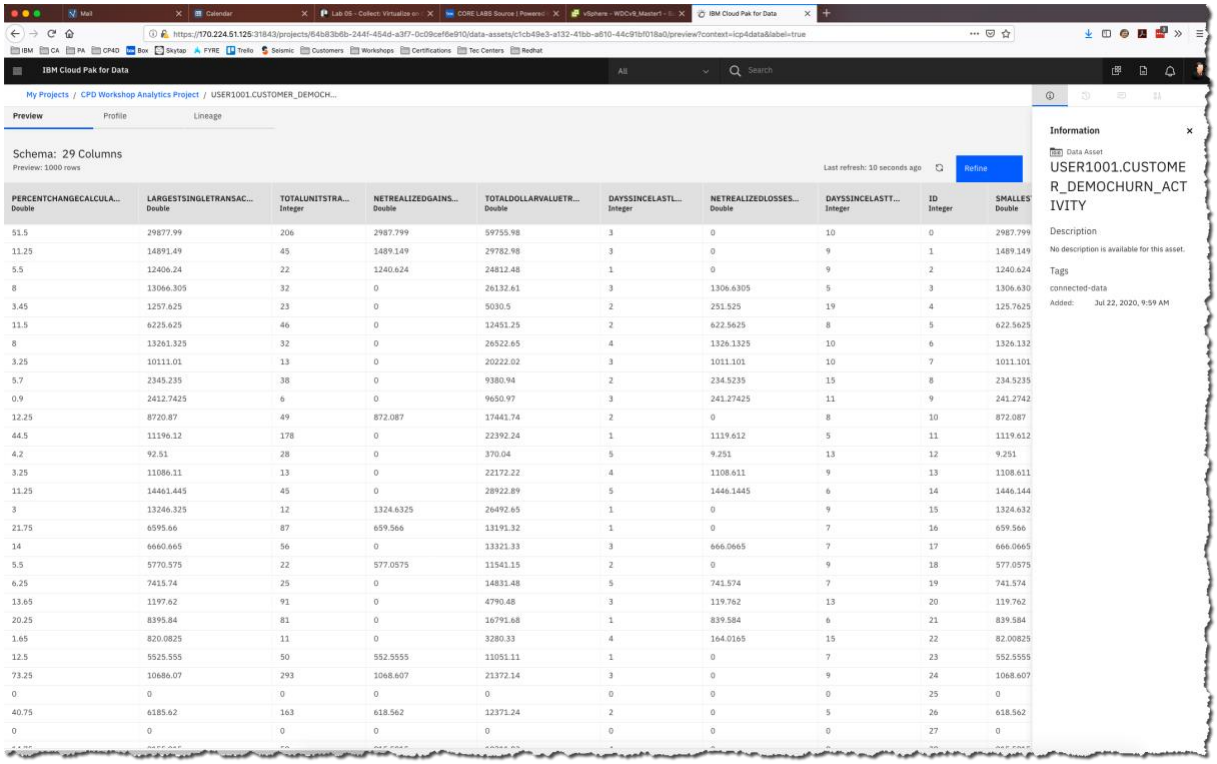
__36. You will be navigated to `My Projects` ⇨ `CPD Workshop Analytics Project` ⇨ `Assets.`

Find and click on your new virtualized view (of the two virtual tables) `USER1001.CUSTOMER_DEMOCHURN_ACTIVITY.`



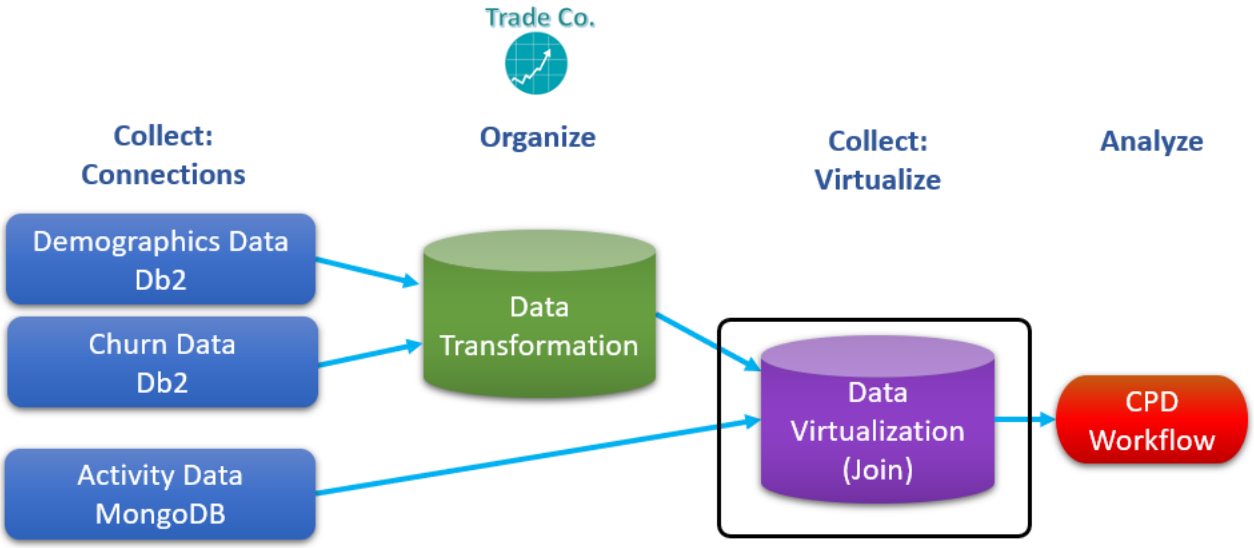
__37. Scroll through the data in the preview mode.

Notice at this point you could refine the data even further. (We will do that in another exercise.)



5.9 Lab conclusion

In this lab, you learned about `Data Virtualization` to complete the `Collect` tasks by creating a virtualized view of the transformed Db2 Demographics and Db2 Churn data, joined with the MongoDB (JSON) Activity data.



 Data Engineer	The SQL query runs on the remote data sources (Db2 and MongoDB) when you call the data set from the project. This capability provides the absolute latest "current state" information from the activity (i.e. transactional data store). If you want to learn even more about data virtualization, see: `(Deeper Dive) Data Virtualization (Caching)` Final note: Cloud Pak for Data refers to joins of individual virtualized tables as "virtualized views."

**** End of Lab 05 - Collect: Virtualize**

Lab by Burt Vialpando and Kent Rubin, IBM

Lab 06 ANALYZE: AUTOAI

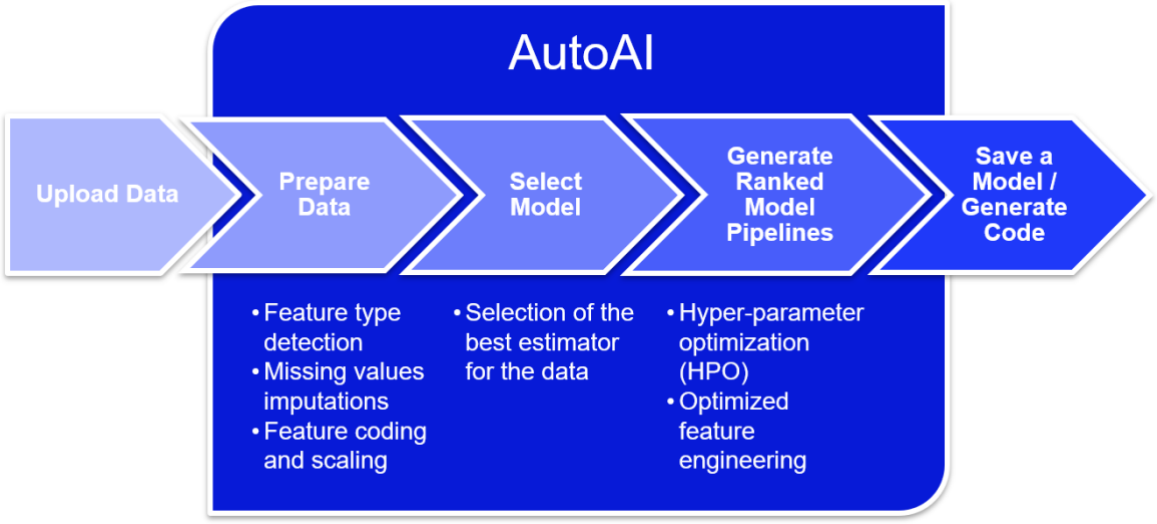
6.1 Lab overview

Analyze is the third phase in the Cloud Pak for Data platform and workflow. This is where data scientists and business analysts can join forces to gain insights from their organization's data. Analyze capabilities can be delivered with a number of different CPD services. Examples:

- AutoAI (included with the Watson Machine Learning service)
- Notebook model creation (included with the Watson Studio service)
- SPSS Modeler
- Decision Optimization
- Cognos Analytics (or Cognos Analytics Embedded)
- IBM Streams (which could be "Collect" or "Analyze" depending on how it is used)
- RStudio model creation...and others

In this lab you will explore AutoAI, which helps simplify the Machine Learning model AI lifecycle by automating the following:

- Data preparation
- Model development
- Feature engineering
- Hyper-parameter optimization



6.2 Persona represented in this lab

The Data Scientist persona is the most likely role to perform the Analyze tasks in this lab, that is, to create a machine learning model with AutoAI that can be deployed and infused into an AI application.

Persona (Role)	Capabilities
 Data Scientist	Data Scientists bring expertise in statistics and the process of building ML/AI models to make predictions and answer key business questions.

6.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Click the desktop icon: `Cloud Pak for Data Web Client.`

__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

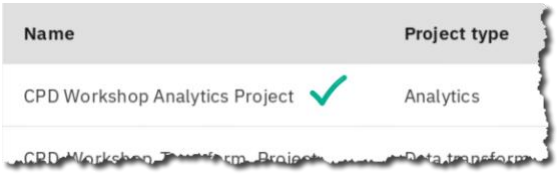
6.4 Setting up the AutoAI experiment

In the lab, you will create an AutoAI experiment that will be used to automatically create the machine learning model that best fits the data to provide the desired outcome. One only needs to provide general guidance, and AutoAI will do the rest of the work.

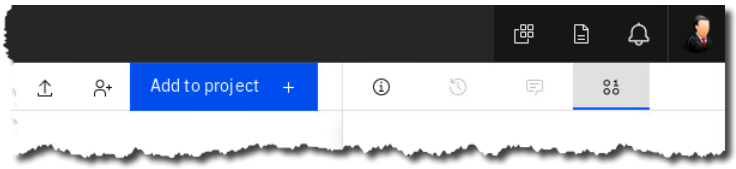
In our scenario, Trade Co. data scientists accelerate their time to value using this powerful tool.

__4. In the CPD web client, click the `Navigation Menu` (“hamburger” icon) ⇨ `Projects.`

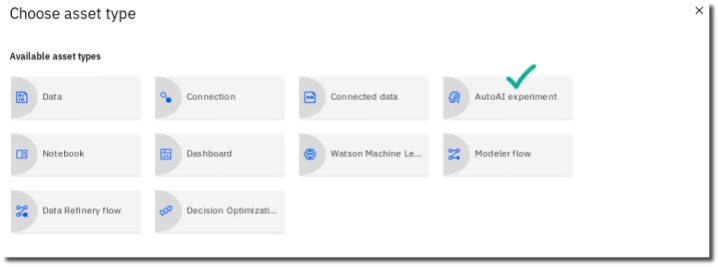
__5. Select the project: `CPD Workshop Analytics Project.`



__6. Once the project is opened, at the top right corner of the screen click on: `Add to project +.`



__7. In the next screen hover over and then click on the tile: `AutoAI experiment.`

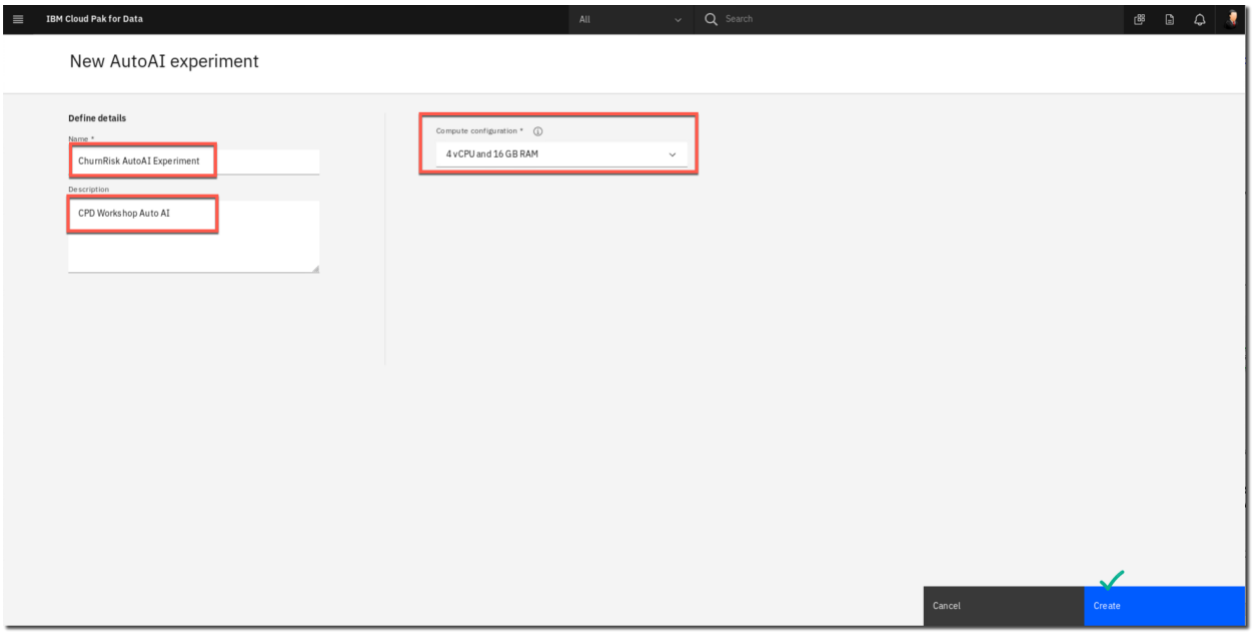


__8. In `Name`, enter `ChurnRisk AutoAI experiment.`

Fill in anything you like in `Description`.

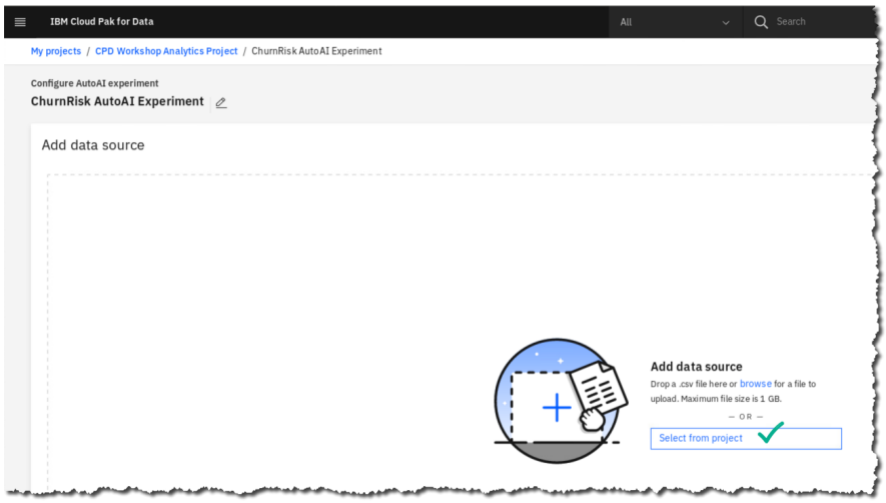
Leave the defaults for `Compute configuration.`

Click `Create.`



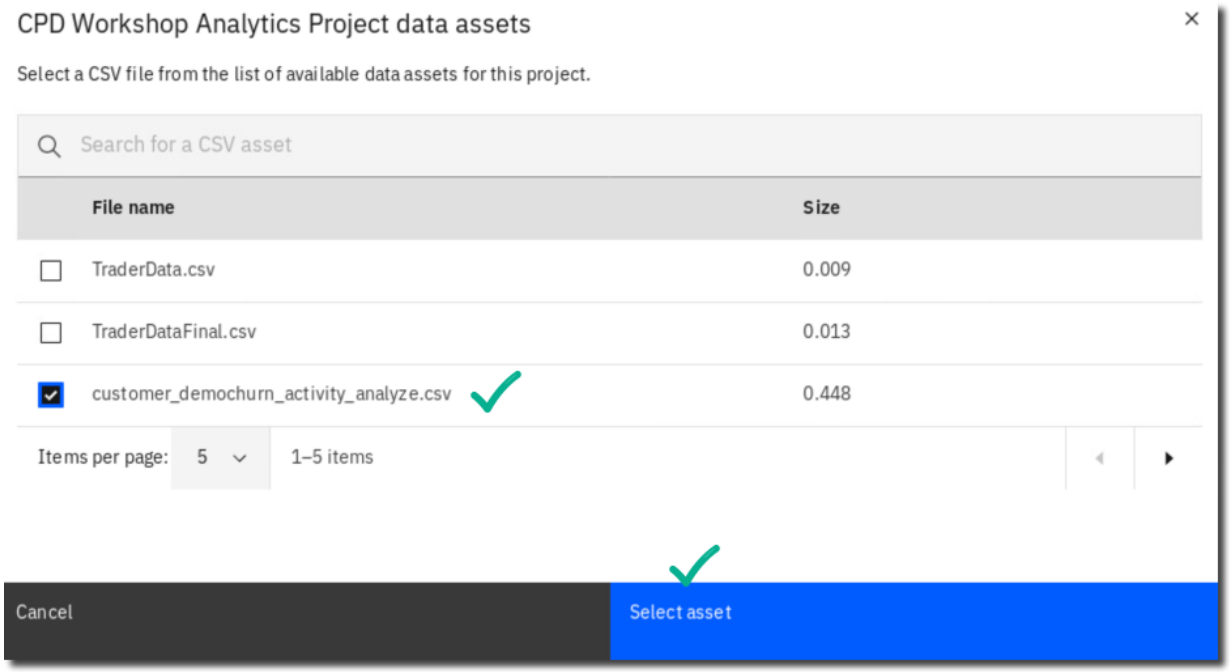
Note: If you have to wait a long time for the spinning circle to complete, simply return to the CPD Workshop Analytics Project and select ChurnRisk AutoAI Experiment from the list of Assets.

__9. In the screen `Add data source`, click `Select from project.`



__10. Select File name: `customer_demochurn_activity_analyze.csv.`

Click `Select asset.`



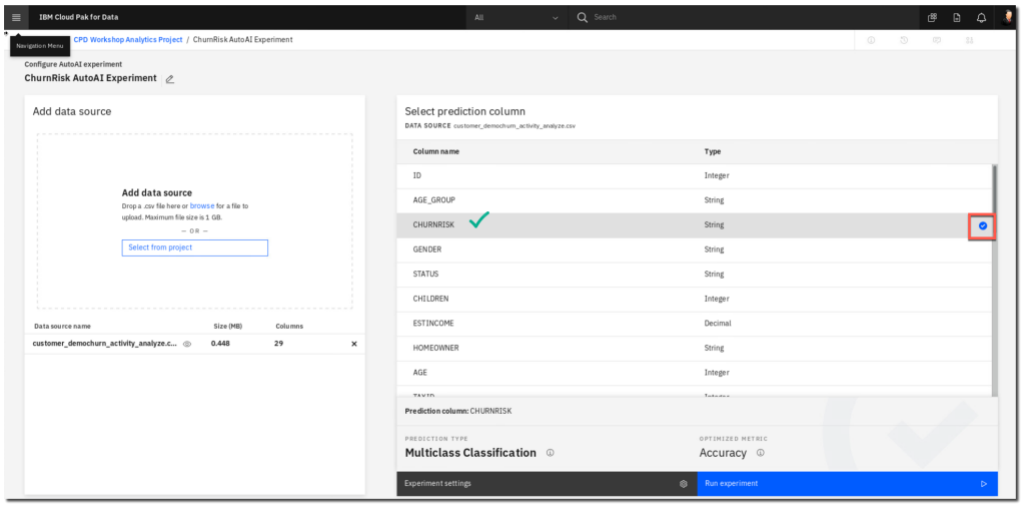
Data Scientist

The CSV file used in this AutoAI experiment is a join of Db2 CUSTOMER_CHURN, Db2 CUSTOMER_DEMOGRAPHICS and MongoDB CUSTOMER_ACTIVITY data.

If you have been doing all the labs so far in this workshop, you would have completed the previous Data Flow Designer and Data Virtualization labs that transformed and joined these tables together as one virtualized view.

Since AutoAI requires a file as input, this virtualized view was exported to a CSV file (`customer_demochurn_activity_analyze.csv`) to be used as input for this lab.

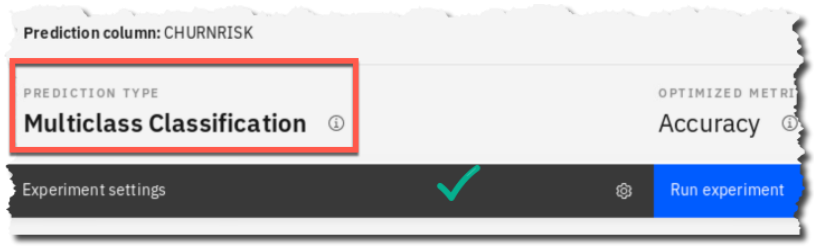
__11. In the screen `Select prediction column`, select `Column name`: `CHURNRISK.`



__12. Notice that the bottom of this screen now fills in when CHURNRISK is select as the Prediction column.

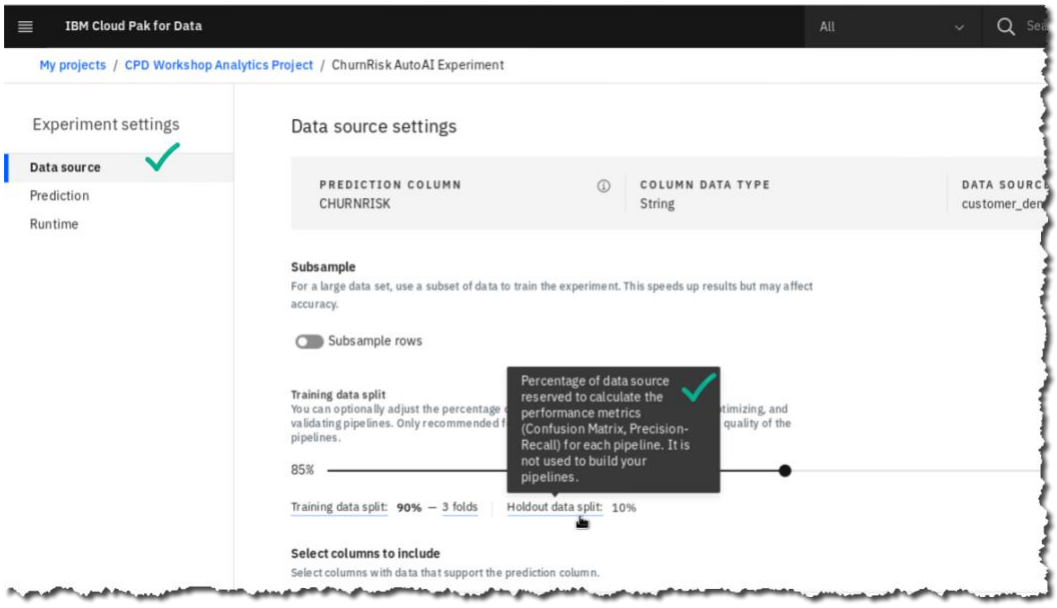
AutoAI has determined for you a PREDICTION TYPE = Multiclass Classification.

Click `Experiment settings.`



__13. The first Experiment setting you can change from the values chosen for you is `Data source`

This allows you to change the `Holdout data split` (for testing vs. training)



If you scroll down you can see that you can optionally choose to select which columns to include in the experiment as well.

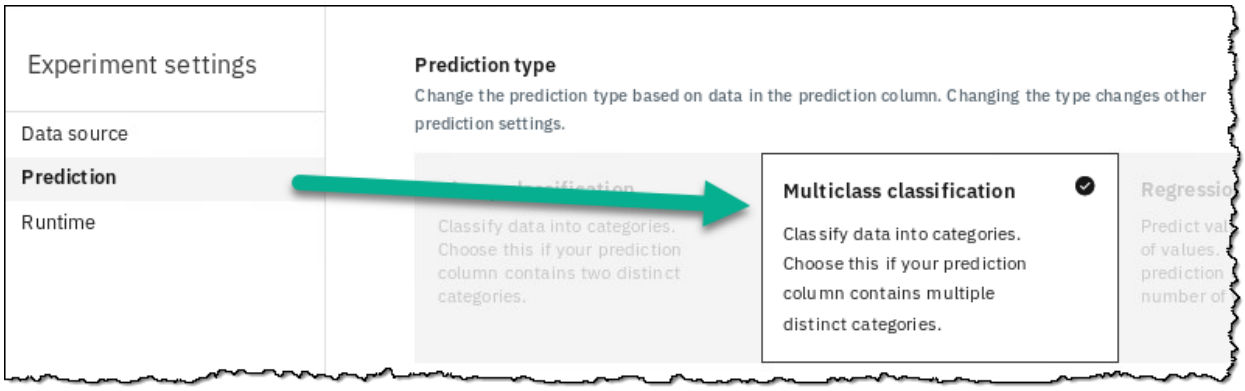
Leave these settings as-is and click on the section: `Prediction.`

__14. The second setting you can change is `Prediction`.

The first section in the Prediction settings is the `Prediction type.`

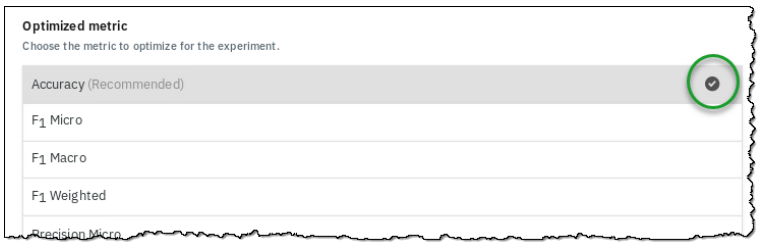
Read what all three Prediction types are best suited for. The one chosen for you was derived from the column `CHURNRISK`, which, by the way, has three values - High, Medium and Low. Thus, AutoAI determined that Multiclass classification is best suited for this data.

Leave these setting as-is .

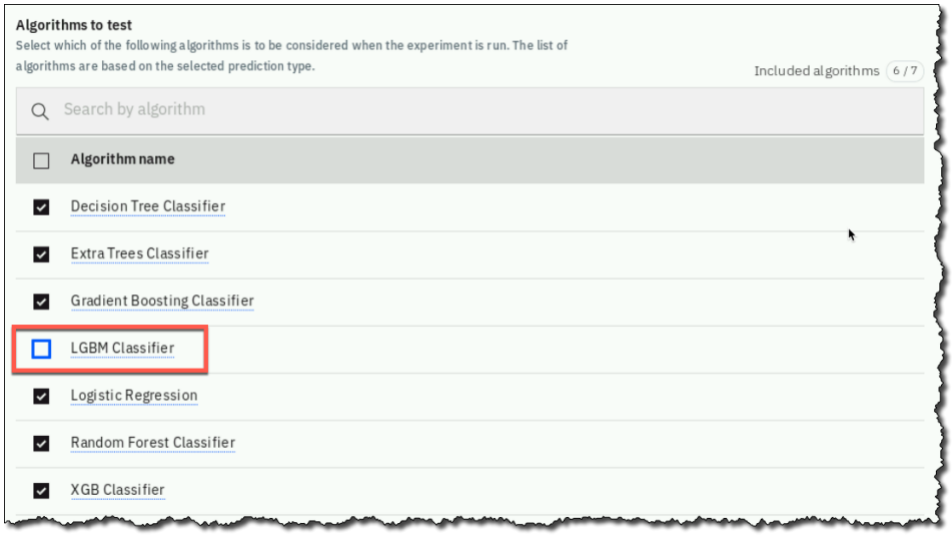


__15. Scroll down to review the next section: `Optimized metric`. Notice that AutoAI chose `Accuracy` for you. Hover over a few of the other metrics you could choose if you so desired.

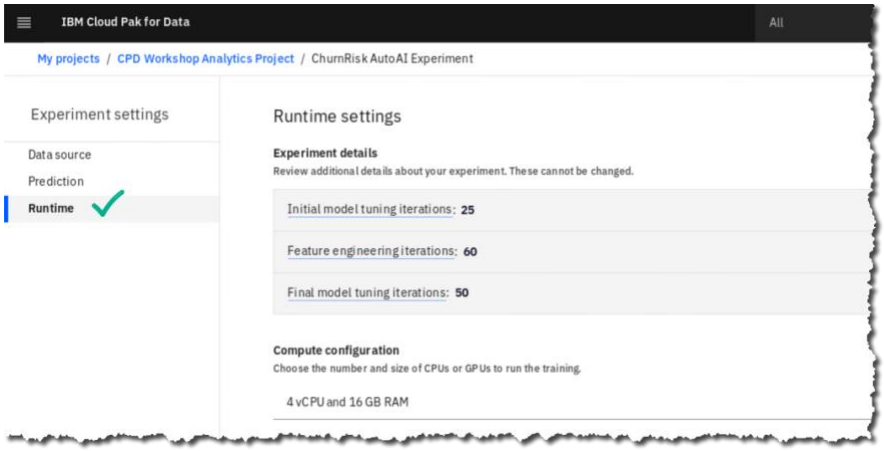
A Data Scientist would best determine if and when to deviate from this recommended metric, but you will leave this choice as-is.



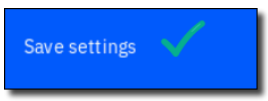
__16. Scroll to review third section which allows you to choose which algorithms to test. `Remove the LGBM Classifier` from the list by deselecting it.



__17. Click the `Runtime` setting to review the last settings for your experiment.

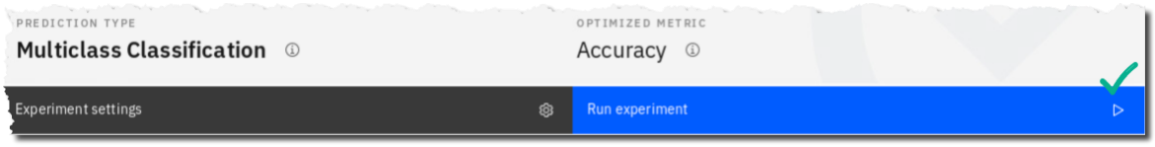


__18. Click `Save settings` to save your changes.

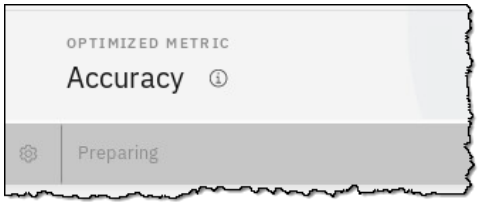


6.5 Running the AutoAI experiment

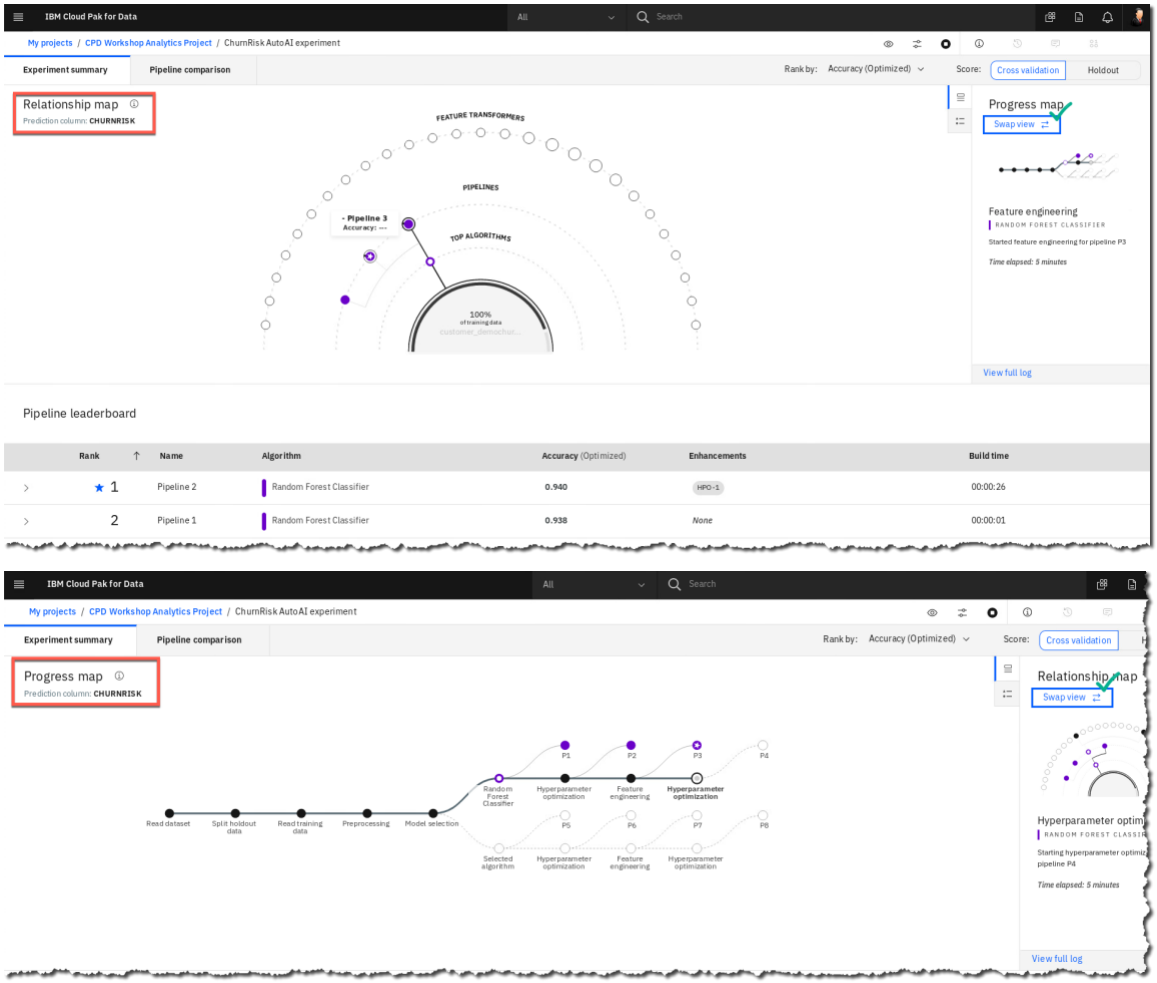
__19. At the bottom of the screen, click `Run experiment.`



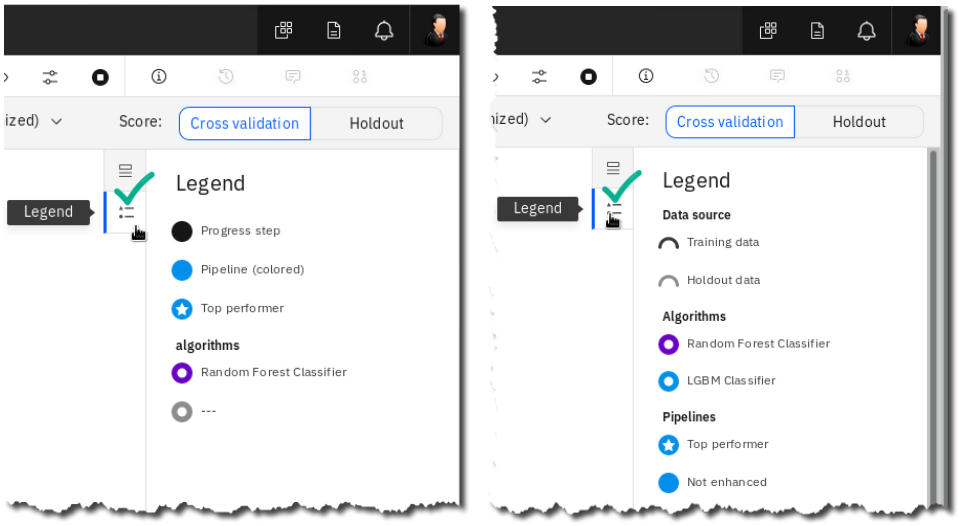
__20. You will first see AutoAI go into a `Preparing` mode… then it will be `Pending`.



__21. Once it is running, you can click on `[Swap view]` to see each of the two different infographics: `Relationship map` and `Progress map.`



__22. Review the `Legend` for each infographic map.



__23. On either infographics map, you can scroll down to see the `Pipeline` leaderboard.

AutoAI begins building several different pipelines (possible solutions for the best model) ranking them as it goes along. You will see the pipelines complete and rank as the process continues.

Pipeline leaderboard

Rank	Name	Algorithm	Accuracy (Optimized)	Enhancements
★ 1	Pipeline 4	Random Forest Classifier	0.950	HPO-1 FE HPO-2
2	Pipeline 8	Decision Tree Classifier	0.944	HPO-1 FE HPO-2
3	Pipeline 3	Random Forest Classifier	0.943	HPO-1 FE
4	Pipeline 2	Random Forest Classifier	0.940	HPO-1
5	Pipeline 7	Decision Tree Classifier	0.939	HPO-1 FE
6	Pipeline 6	Decision Tree Classifier	0.938	HPO-1
7	Pipeline 1	Random Forest Classifier	0.938	None
8	Pipeline 5	Decision Tree Classifier	0.891	None

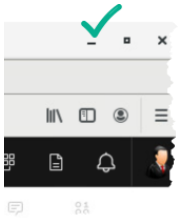
__24. As AutoAI does its work, watch it flow through its various steps:

- **Read dataset:** Reads the data set you provided for the experiment.
- **Split holdout data:** Splits the data between testing and training.
- **Read training data:** Reads training data to prepare for preprocessing.
- **Preprocessing:** Most data sets contain different data formats and missing values, but standard ML algorithms work with numbers and no missing values. AutoAI applies various algorithms (estimators) to analyze, clean, and prepare your raw data for machine learning. It automatically detects and categorizes features based on data type, such as categorical or numerical. It determines the best combination of strategies for missing value imputation, feature encoding, and feature scaling for your data.
- **Model selection:** AutoAI uses a novel approach that enables testing and ranking candidate algorithms against small subsets of the data, gradually increasing the size of the subset for the most promising algorithms to arrive at the best match by ranking large numbers of candidate algorithms. This approach saves time without sacrificing performance.
- **Selected estimator:** Shows the estimator chosen from the model selection step.
- **Hyperparameter optimization:** Refines the best performing model pipelines by using a novel hyper-parameter optimization algorithm optimized for costly function evaluations such as model training and scoring that are typical in machine learning. This approach enables fast convergence to a good solution despite long evaluation times of each iteration.
- **Feature engineering:** Attempts to transform the raw data into the combination of features that best represents the problem to achieve the most accurate prediction. AutoAI uses a novel approach that explores various feature construction choices in a structured, non-exhaustive manner, while progressively maximizing model accuracy using reinforcement learning. This results in an optimized sequence of transformations for the data that best match the algorithms of the model selection step.

6.6 Running a Notebook

While the AutoAI experiment is running, **open another** CPD web client to perform a parallel exercise.

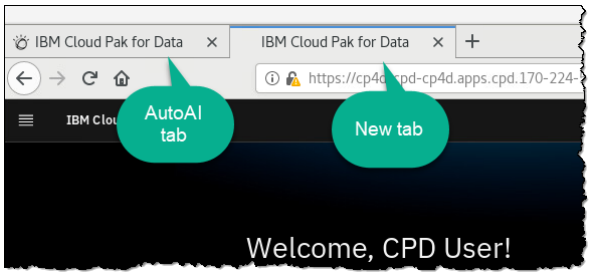
__25. Minimize the CPD web client browser to be able to get to the desktop.



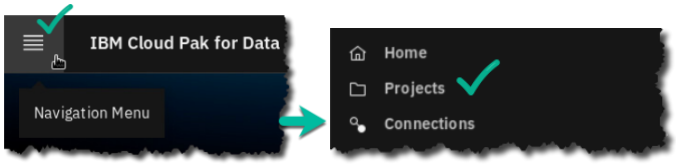
__26. Double click the desktop icon: `Cloud Pak for Data Web Client.`

__27. This will open a second CPD Web client browser tab.

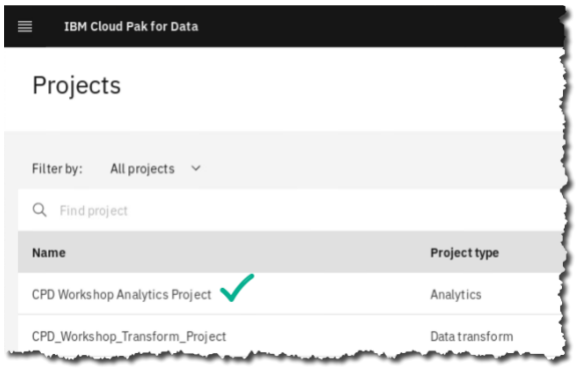
The AutoAI session is running in the first tab **so don't close it!**



__28. In the CPD web client, click the `Navigation Menu` ⇨ `Projects.`

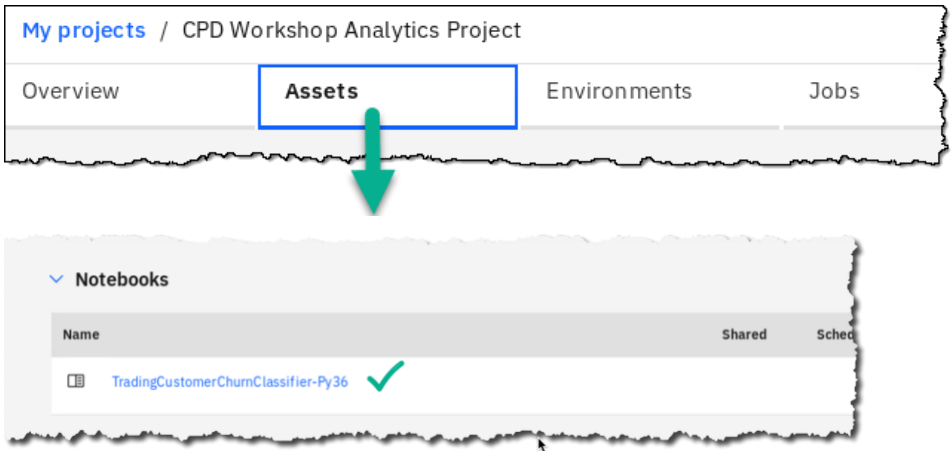


__29. Select the project: `CPD Workshop Analytics Project`.

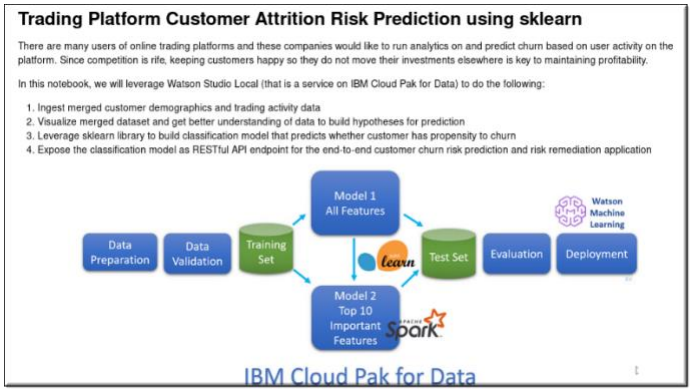


__30. In the section `Assets`, scroll down to find `Notebooks.`

Click `TradingCustomerChurnClassifier-Py36.`

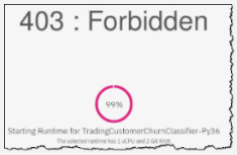


__31. You will be presented with the opened notebook.

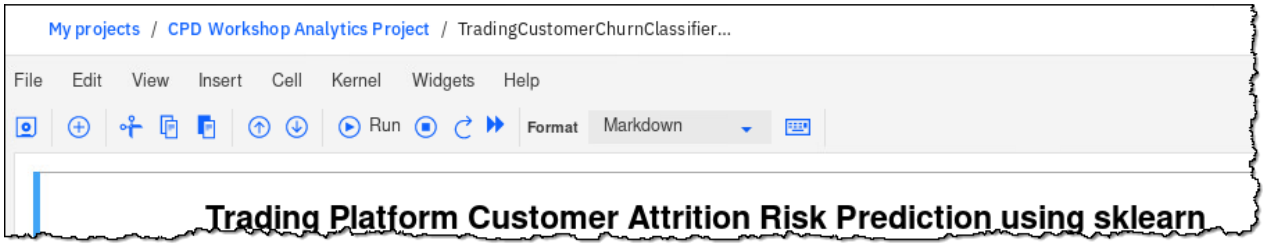


__32. Click the `Edit` (pencil) `icon` to put the Notebook in edit mode.

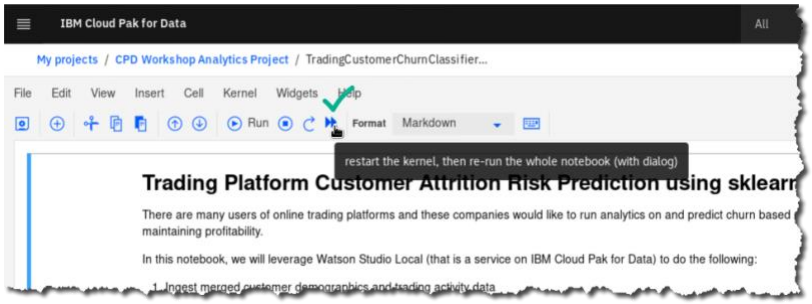
(Note: If you are returning to the Notebook, it may already be in edit mode.)

 Data Scientist	Note: if you received an error like this “403: forbidden.”  Fix the problem by leaving this screen and going to `My Instances` ⇨ `Environments` and then delete the Runtime environment for Python 3.6 that is currently running. Try opening the notebook again after returning to the project. 

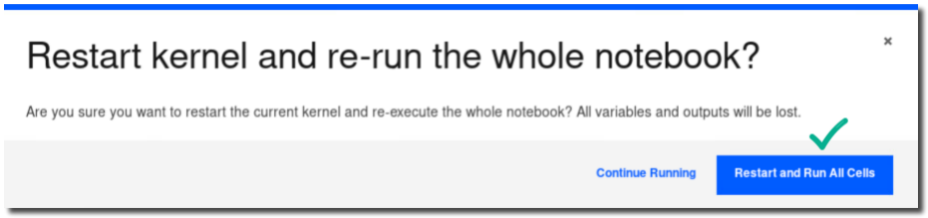
__33. The Notebook will start a runtime and present a screen with the top left looking like this:



__34. Click on the `Fast Forward` (double arrow) `icon` to re-run the whole notebook.

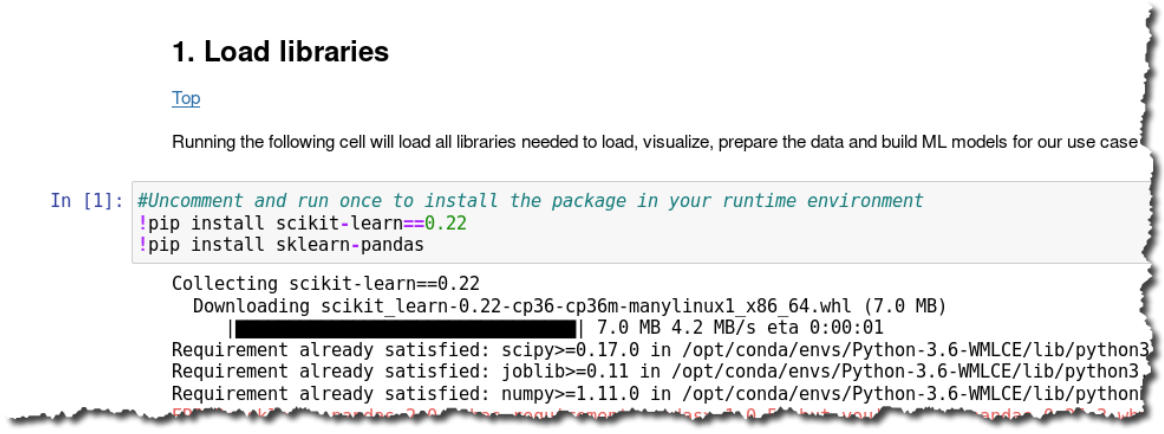


__35. Click `Restart and Run All Cells.`



__36. The notebook will now start running each cell in sequential order. A number will be placed next to each cell after it has executed.

Scroll up to the first cells under `1. Load libraries`.



 Data Scientist	Note: You can run the notebook cell by cell or all at once. Either way will give you the same result. Any cell that has not yet run is indicated like this: `[*]`

__37. Scroll to the 4th cell (under `2. Load data example`) to see the input file for this notebook. Notice it is the same input file from the same project that AutoAI is using.

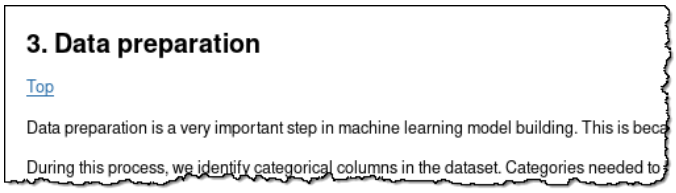
In [4]:
```
df_churn_pd = pd.read_csv('/project_data/data_asset/customer_demochurn_activity_analyze.csv')
df_churn_pd.head()
```

Out[4]:

	ID	AGE_GROUP	CHURNRISK	GENDER	STATUS	CHILDREN	ESTINCOME	HOMEOWNER	AGE	TAXID	...	LATITUDE
0	0	Young adult	Low	F	S	1	38000.00	N	24	147889187	...	NaN
1	1	Adult	Low	M	M	2	29616.00	N	49	113772166	...	38.687261
2	2	Adult	Low	M	M	0	19732.80	N	51	132420919	...	NaN
3	3	Adult	High	M	S	2	96.33	N	56	700548452	...	32.531971
4	4	Young adult	High	F	M	2	52004.80	N	25	141013706	...	33.593192

Scroll through this section of the notebook to see various visualizations of the data.

__38. Review the cells in `3. Data preparation.`



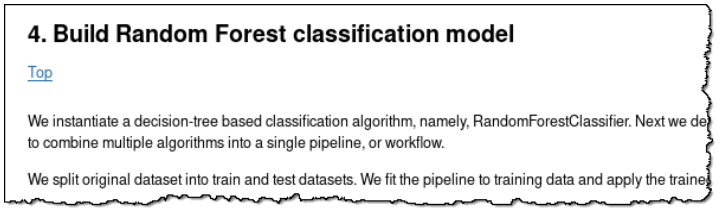
3. Data preparation

Top

Data preparation is a very important step in machine learning model building. This is beca

During this process, we identify categorical columns in the dataset. Categories needed to

__39. The next section is self-explanatory: `4. Build Random Forest classification model.`



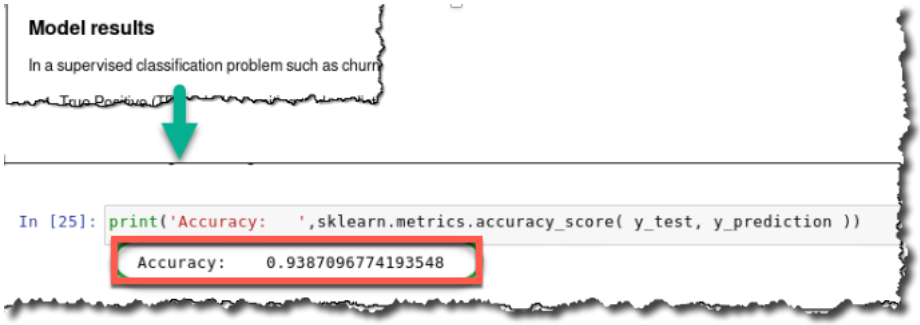
4. Build Random Forest classification model

Top

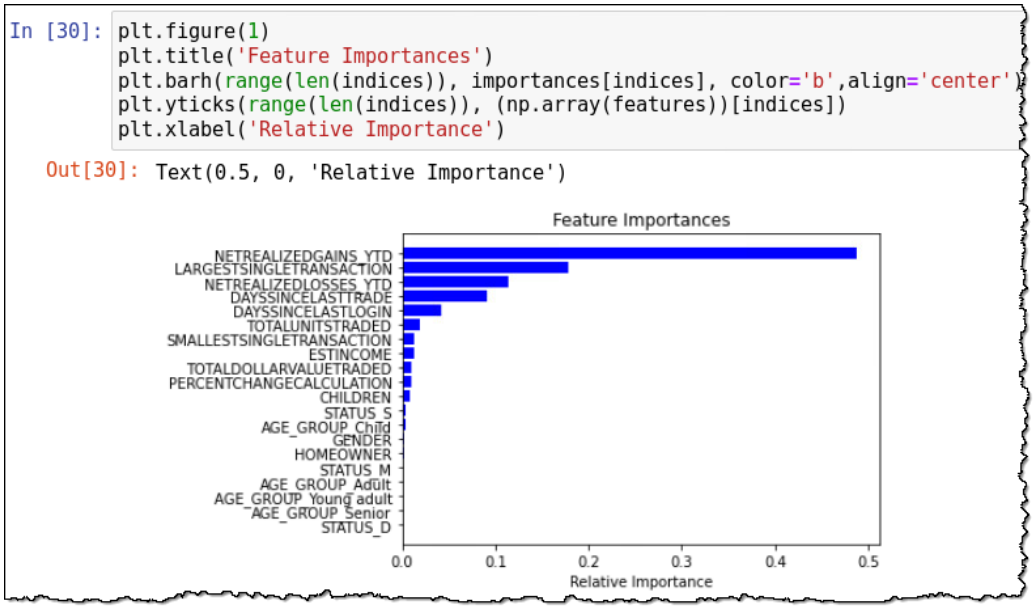
We instantiate a decision-tree based classification algorithm, namely, RandomForestClassifier. Next we de to combine multiple algorithms into a single pipeline, or workflow.

We split original dataset into train and test datasets. We fit the pipeline to training data and apply the traine

__40. Scroll to find `Model Results`, then find the `Accuracy.`

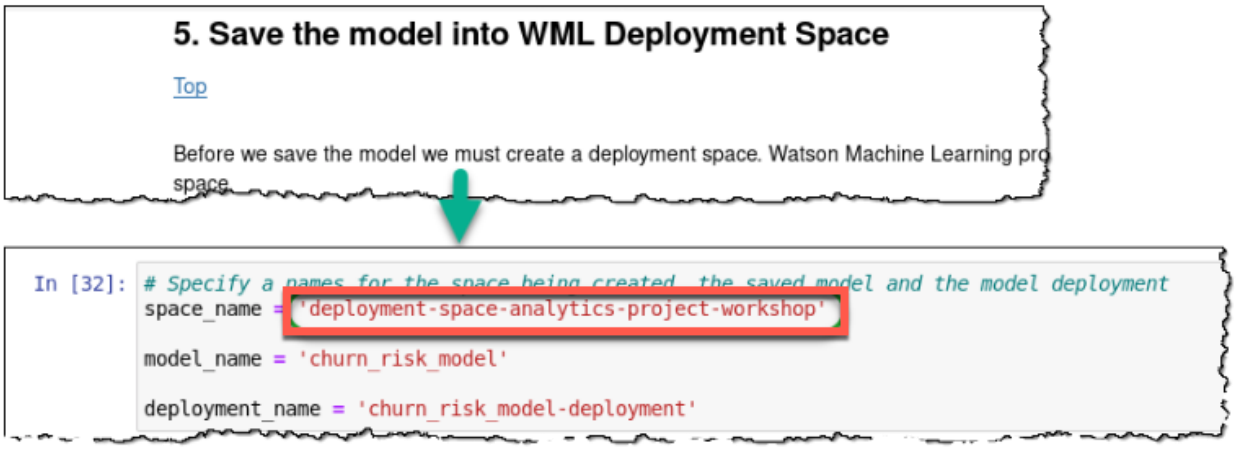


__41. Cell [30] shows the “Feature Importances,” that is, the data columns that affected the model the most

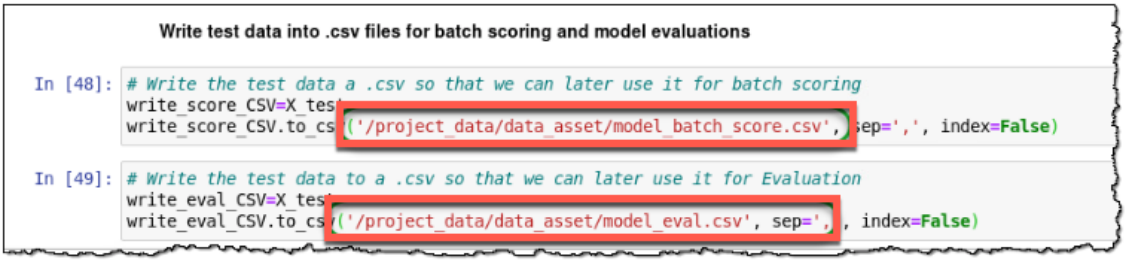


__42. Finally, review the name listed in `5. Save the model into WML Deployment Space.`

We will be referring to this Deployment Space in a later lab.

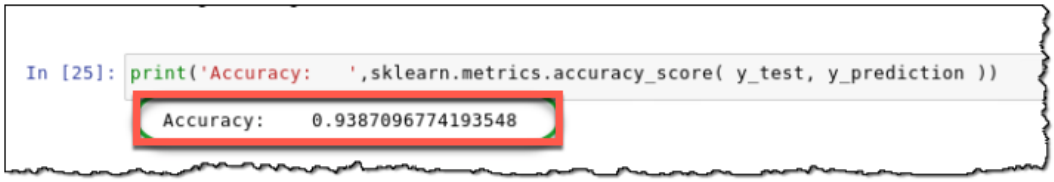


__43. Also note at the very end of the notebook in the last two cells that two files are created for batch scoring and evaluation.



Your notebook has NOT finished until you see that the last two code cells (above) have a number from the run.

__44. One key data point to take particular note of from this notebook run is found in cell 25, which is the accuracy of the Random Forest model created by this notebook.



Data Scientist

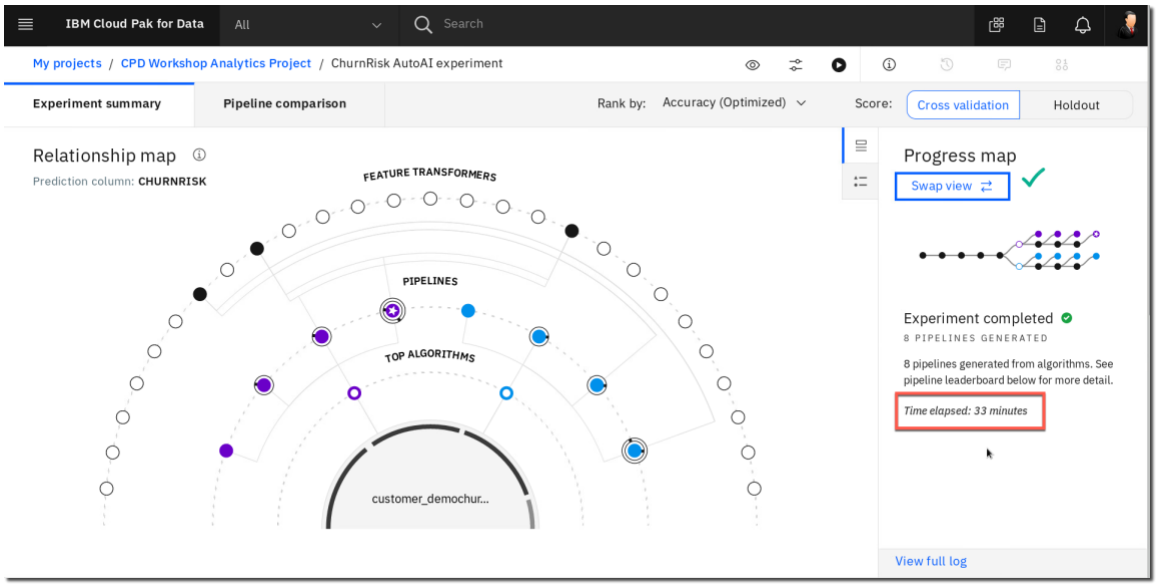
Ten Reasons Why I Like my Jupyter Notebook

1. **All in one place**: The Jupyter Notebook is a web-based interactive environment that combines code, rich text, images, videos, animations, mathematical equations, plots, maps, interactive figures and widgets, and graphical user interfaces, into a single document.
2. **Easy to share**: Notebooks are saved as structured text files (JSON format), which makes them easily shareable.
3. **Easy to convert**: Jupyter comes with a special tool, nbconvert, which converts notebooks to other formats such as HTML and PDF.
4. **Language independent**: The architecture of Jupyter is language independent. The decoupling between the client and kernel makes it possible to write kernels in any language.
5. **Easy to create kernel wrappers**: Jupyter brings a lightweight interface for kernel languages that can be wrapped in Python. Wrapper kernels can implement optional methods, notably for code completion and code inspection.
6. **Easy to customize**: Jupyter's interface can be used to create an entirely customized experience in the Jupyter Notebook (or another client application such as the console).
7. **Extensions with custom magic commands**: Create IPython extensions with custom magic commands to make interactive computing even easier. Many third-party extensions and magic commands exist, for example, the %%cython magic that allows one to write Cython code directly in a notebook.
8. **Stress-free Reproducible experiments**: Jupyter notebooks can help you conduct efficient and reproducible interactive computing experiments with ease. It lets you keep a detailed record of your work. Also, the ease of use of the Jupyter Notebook means that you don't have to worry about reproducibility; just do all of your interactive work in notebooks, put them under version control, and commit regularly. Don't forget to refactor your code into independent reusable components.
9. **Effective teaching-cum-learning tool**: The Jupyter Notebook is not only a tool for scientific research and data analysis but also a great tool for teaching.
10. **Interactive code and data exploration:** The ipywidgets package provides many common user interface controls for exploring code and data interactively.

6.7 Reviewing the AutoAI results

__45. Return to your first CPD Web client browser tab.

If the AutoAI experiment has not completed yet, let it finish. You can tell if it is finished by looking at the `Relationship map` view.



Data Scientist	Note: this experiment may take 25 minutes or more to complete, but this is not normal. We have turned off AVX/AVX2 processor support for this workshop so that the Cloud Pak for Data workshop image can run on servers that do not have AVX nor AVX2 processors. If we did not do that and the workshop was run on servers without AVX/AVX2 processors, the experiment would fail. So, we took this route to make sure the AutoAI experiment would run under any circumstances on any server, with or without AVX/AVX2 processors. **That said, with a CPD cluster on AVX/AVX2 supported processors this experiment completes on average, in around 3 minutes.** Servers with AVX/AVX2 processors were released in 2011 and are mostly ubiquitous. Your organization probably has them.

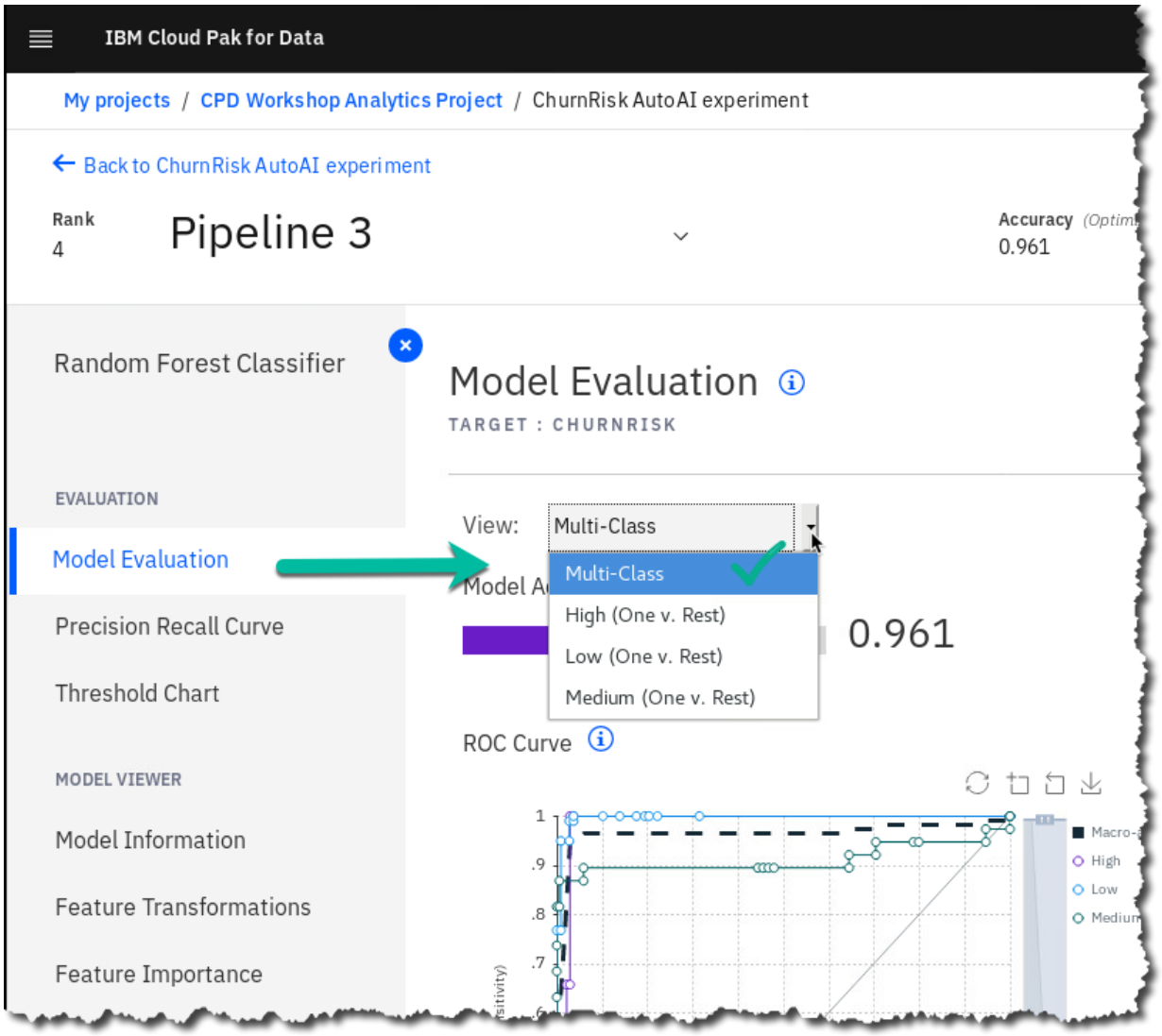
Data Scientist	Note: the results from your AutoAI experiment may vary from the illustrations in this workbook. This is especially true for Feature Transformations.

__46. AutoAI chooses the best model from the various pipeline leaderboard options as Rank #1. You can scroll down in either map infographic screen to see the leaderboard.

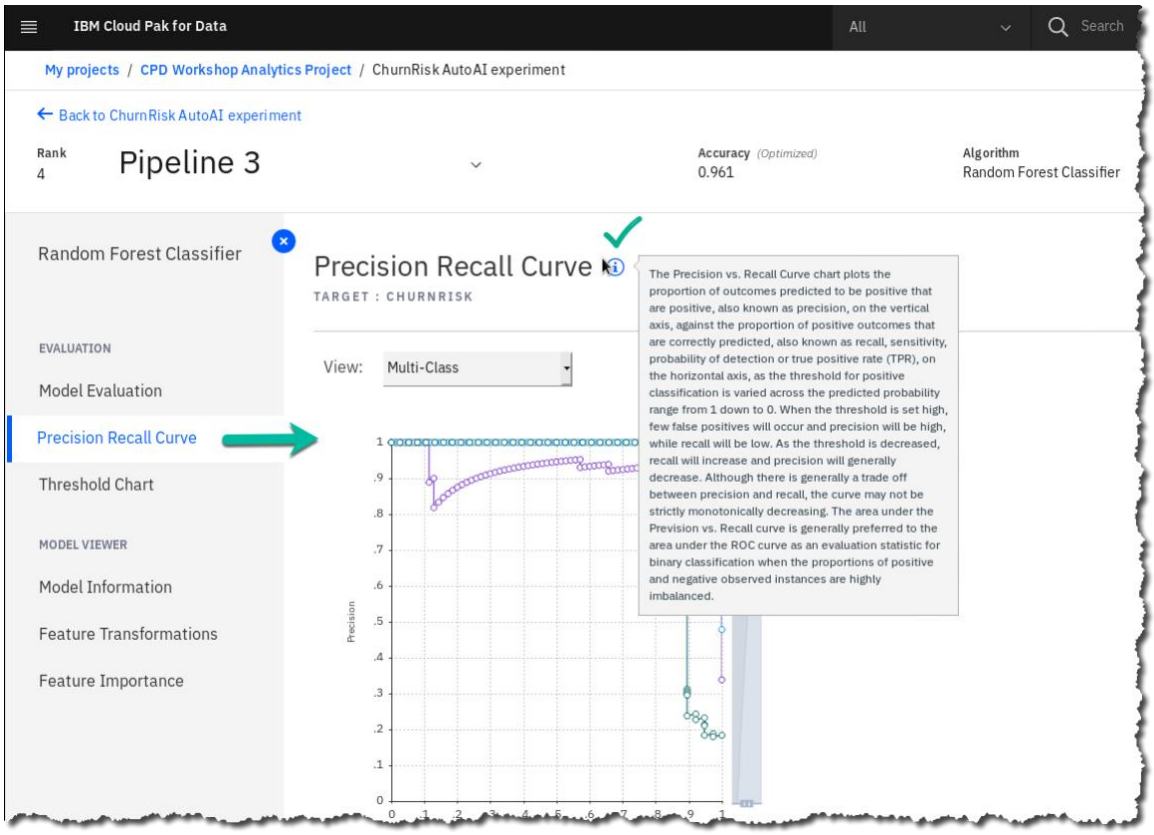
Notice the Algorithm chose (Random Forest classifier) and the most accurate result.

Click on the #1 ranked `Pipeline 4`.

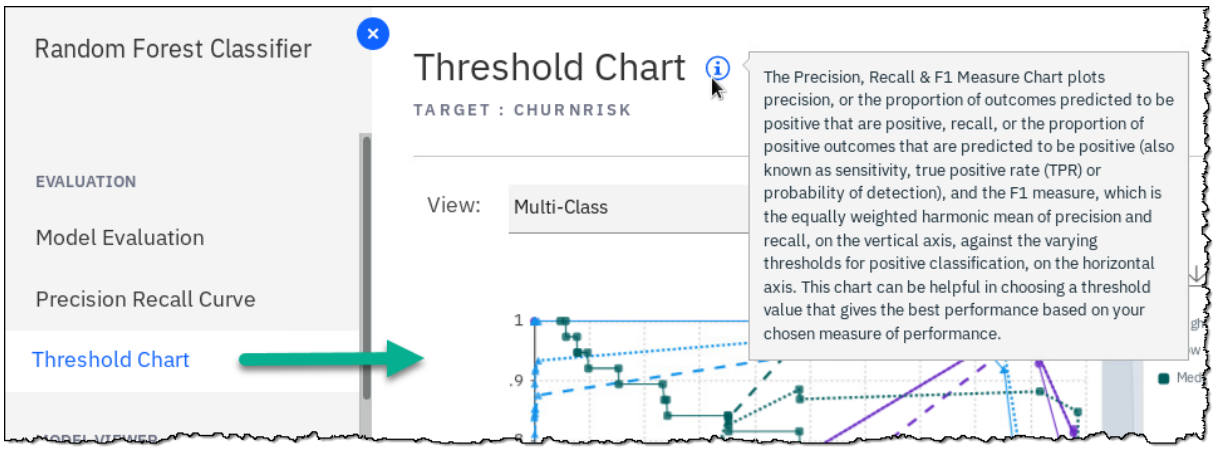
__47. `Model Evaluation` shows the various evaluation accuracy figures. Click through them but make sure to return to the original view - Multi-Class.



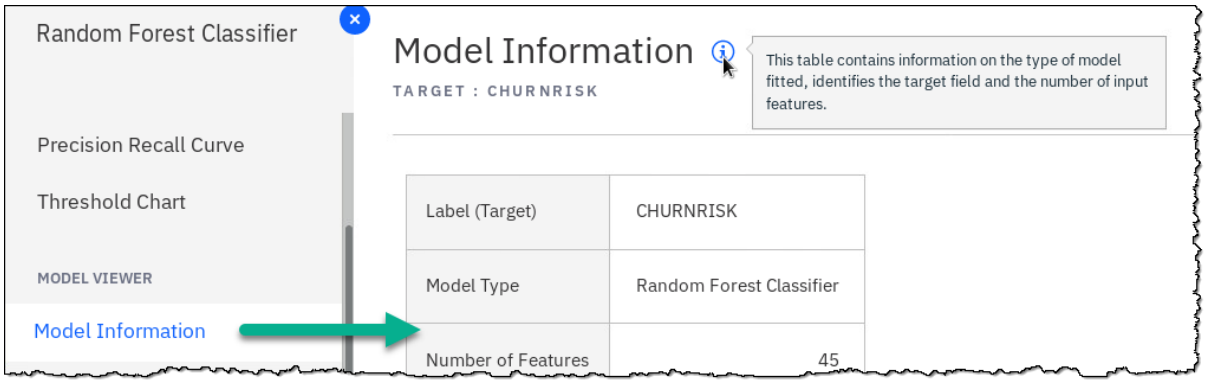
__48. `Precision Recall Curve` Shows the tradeoff between precision and recall.



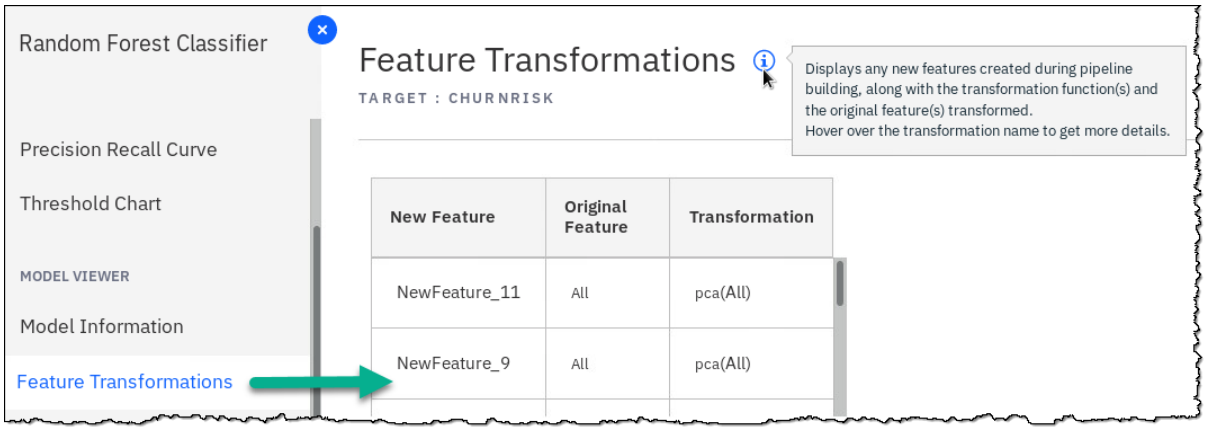
__49. `Threshold Chart` helps choose a threshold for best performance based on the chosen measure of performance.



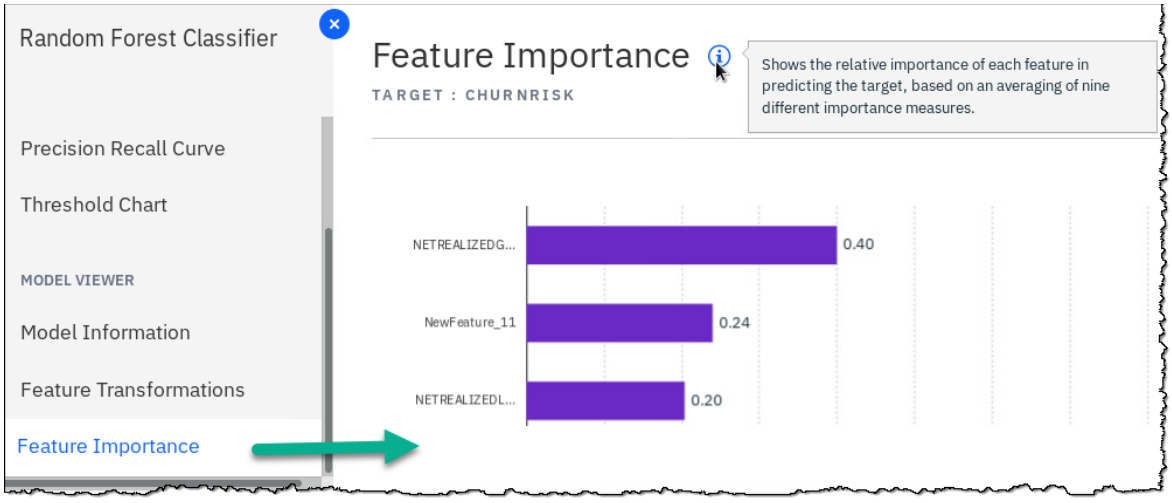
__50. `Model Information` gives info on type of model fitted.



__51. `Feature Transformations` displays new features created by AutoAI. This is a powerful capability of AutoAI because this is not something a Data Scientist may intuitively do themselves in a notebook. (Note: your results may vary from this screen shot.)



__52. `Feature Importance` displays the relative importance of the feature in predicting the target.

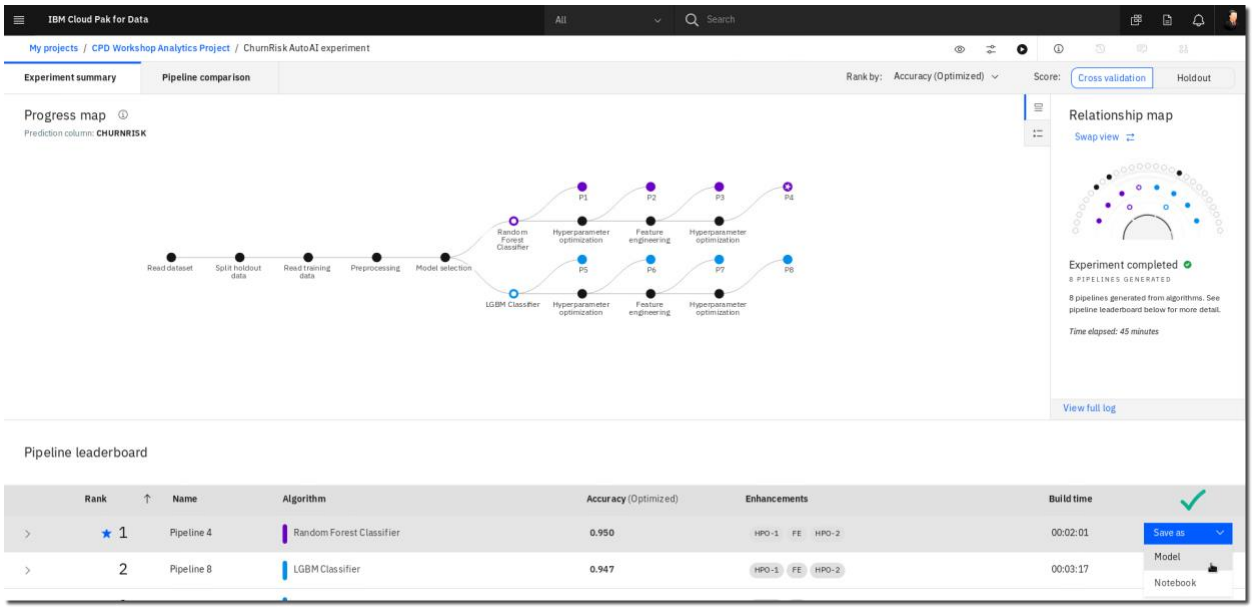


6.8 Saving the model

AutoAI gives you two options to save the model you decide is the best one for your application. You can either save the model as a model or save the model as a Notebook.

6.8.1 Saving the model as a model

__53. Hover your mouse over the right side of your Top Rank model, click `Save as` ⇨ `Model.`

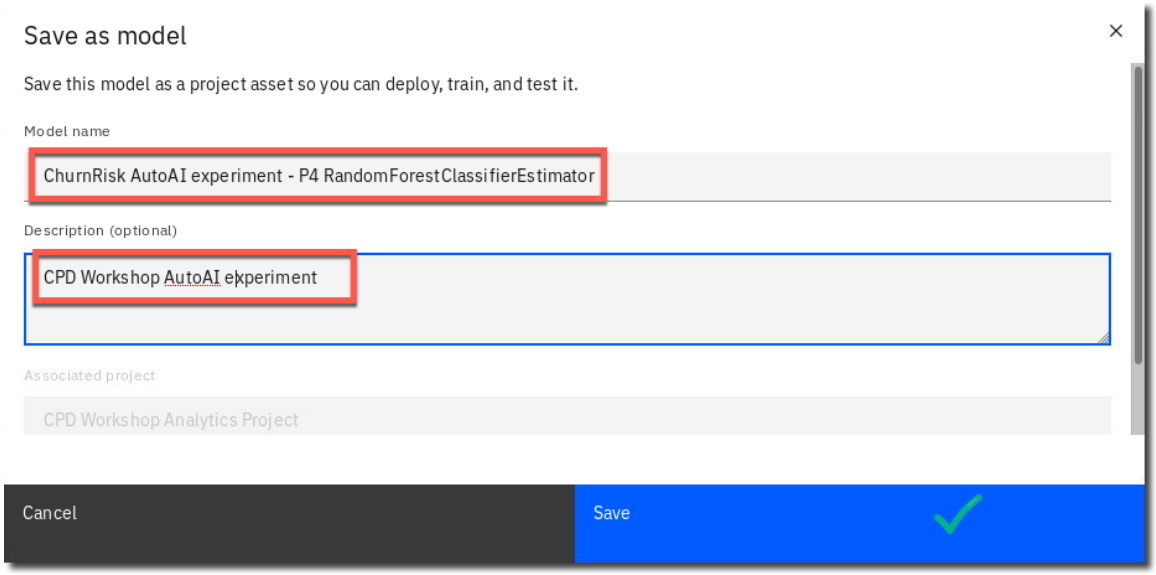


__54. In the screen *Save as model*:

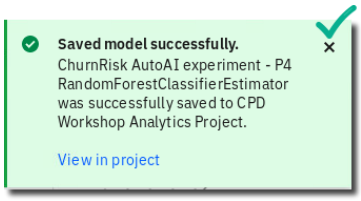
Model name: keep it as is

Description: `CPD Workshop AutoAI experiment`

Click `Save.`

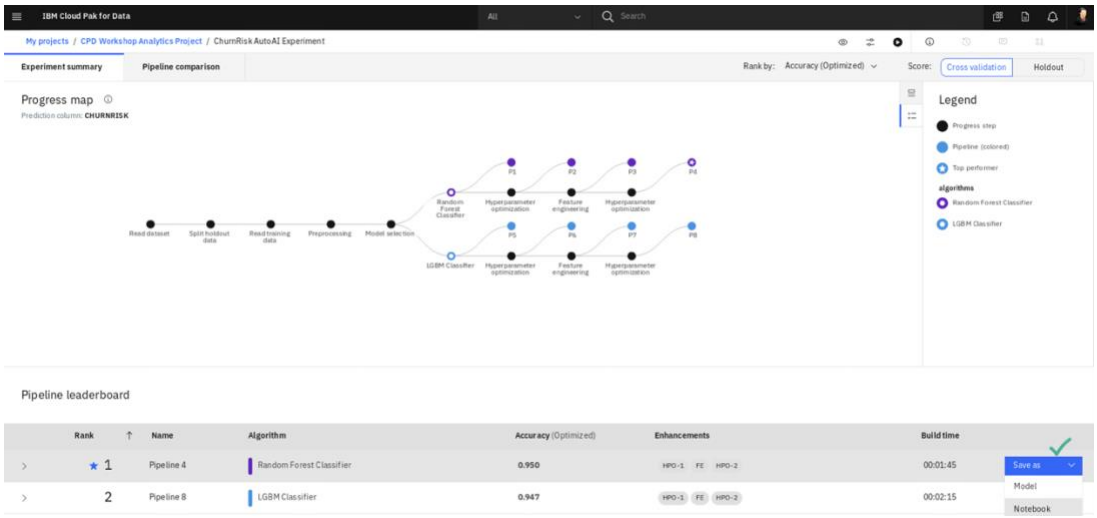


__55. Exit the confirmation pop up by clicking `x`.



6.8.2 Saving the model as a notebook

__56. On the top right corner of the screen, click `Save as` ⇨ `Notebook.`

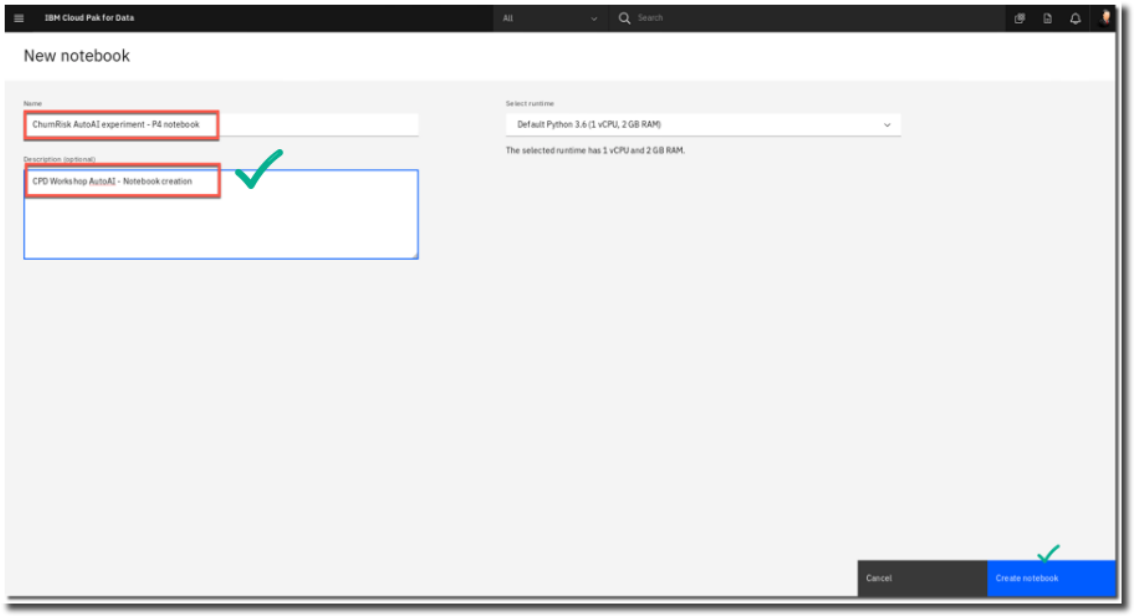


__57. In the screen *New notebook*:

Name: keep it as is

Description: `CPD Workshop AutoAI – Notebook creation`

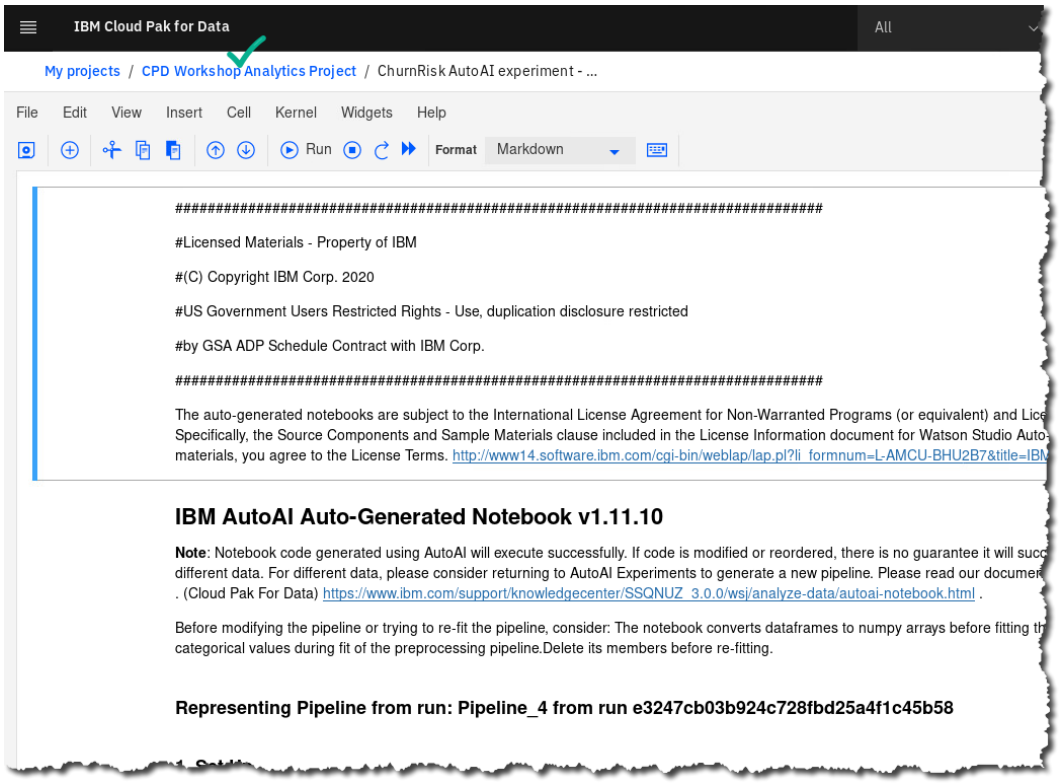
Click `Create notebook.`



__58. You will be taken to the new notebook in edit mode in the project.

Take a moment to examine the new auto-generated notebook you just created.

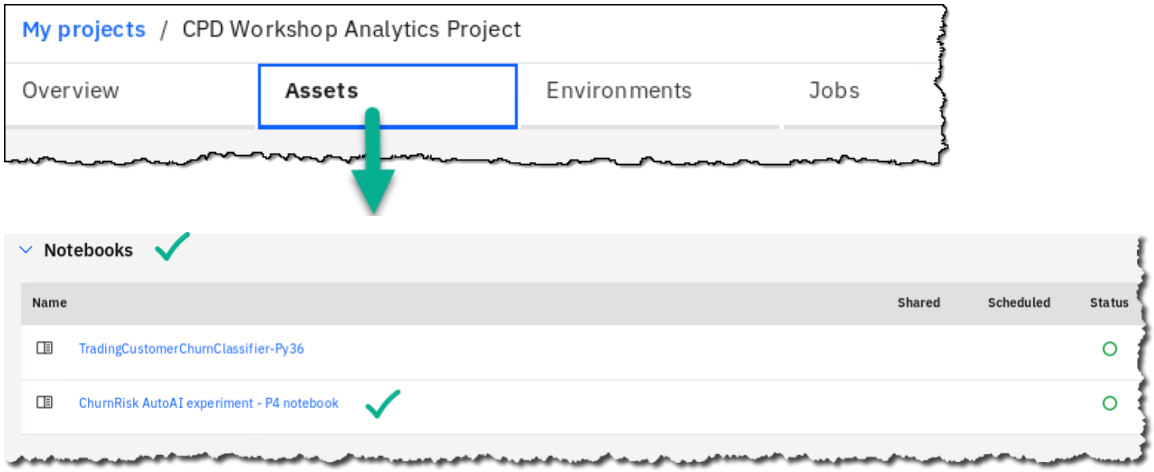
Exit the notebook by clicking `CPD Workshop Analytics Project` on the breadcrumb trail.



6.8.3 Reviewing the project

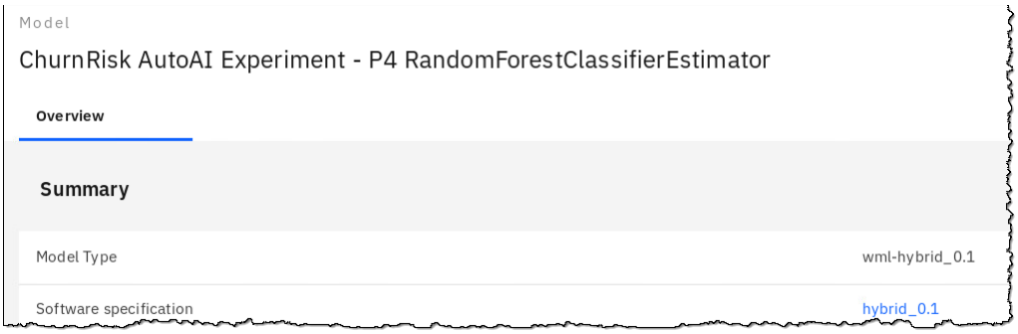
__59. You should be back in the project and under the Assets tab. (If you are not, navigate there now.)

Scroll down to find the Notebooks section. There you will find the AutoAI generated notebook you were just in.

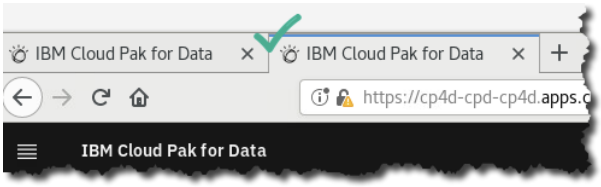


__60. Scroll down farther to find the model you created from the AutoAI save feature. Select the `ChurnRisk AutoAI experiment – P4 RandomForestClassifierEstimator.`

__61. Review your new AutoAI generated model.



__62. Close any extra browser tabs you may have open, leaving only one.



6.9 Lab conclusion

The AutoAI generated model had greater accuracy than the one built by the notebook. AutoAI can test many variations of a model in minutes that a Data Scientist would take days, weeks or even months to try out manually using a notebook approach. And AutoAI does not require any coding at all, making it an available solution to more people.

Data Scientists will always love their notebooks for the reasons listed in this lab, but AutoAI is an industry unique IBM differentiator that will enable the Data Scientist to be more productive. A data scientist can accelerate model creation using AutoAI and then tweak it using a Jupyter notebook if desired. In minutes AutoAI generates this code that would normally take hours or days to create manually.

To fulfill the promise of AI, organizations are tackling skill-set gaps, deployment, and governance processes today. In particular, businesses are seeking an alternative where citizen data scientists can quickly get started, while expert data scientists can speed experimentation time from weeks and months to minutes and hours.

Both groups need a multimodal data science and AI environment where data and analytics specialists collaborate with other experts and optimize model performance end-to-end.

AutoAI in the Cloud Pak for Data Watson Machine Learning service solves this requirement.

Read here how AutoAI won the 2019 Alconics award for *"Best Innovation in Intelligent Automation"*: http://ibm.biz/The-Alconics-Awards

Award for IBM's AutoAI

**** End of Lab 06 - Analyze: AutoAI**

Lab by Burt Vialpando, Kent Rubin, Anjali Shah and Sidney Phoon - IBM

Lab 07 DEPLOY

7.1 Lab overview

Note: This lab requires that you have completed Lab 06 Analyze: AutoAI.

In the previous lab, you created a model from both a notebook and from AutoAI. You will learn how to Deploy these models in this lab.

In our scenario, Trade Co. uses these steps to deploy their machine learning models into production.

7.2 Persona represented in this lab

The Developer persona is the likely role to perform the various Deploy tasks in this lab. However, the Data Scientist persona could perform these tasks as well.

Persona (Role)	Capabilities
 Developer	Developers create and maintain the end-user applications that utilize the output from all the other personas on the CPD platform.

7.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Double-click the desktop icon: Cloud Pak for Data Web Client.

__3. The CPD web client GUI displays as shown. Use cpduser and cpdaccess for the *Username* and *Password* and click Sign in.

7.4 Reviewing the notebook deployment space

Think of model deployment as the equivalent of writing a self-service application that takes the model and makes it available through a REST API interface. Application developers access and consume the model through the same interface. While this is a manual process in most organizations, Cloud Pak for Data can automate deploying and maintaining models without writing a single line of code.

CPD eliminates the need for to do the following:

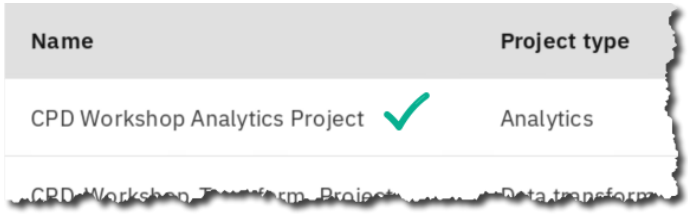
- Writing code to perform the above capability and using a runtime to deploy it.
- Creating a runtime on bare metal machines that require OS installation, network, storage, etc.
- Creating a runtime on a virtual machine in VMware on Intel, or IBM POWER VM® on a POWER platform.
- Creating a runtime in Docker or CRI-O requires someone to build the image and deploy it on one of the above platforms.

Each of the above requires manpower and machine resources. Using CPD, you can bypass this and quickly harvest insight from your data in a repetitive manner by integrating it with your CI/CD pipeline.

7.4.1 Associating a deployment space

__4. In the CPD web client, click the Navigation Menu a Projects.

__5. Select the project: CPD Workshop Analytics Project.



__6. Click section Settings.

__7. Scroll down to find Associated deployment space,

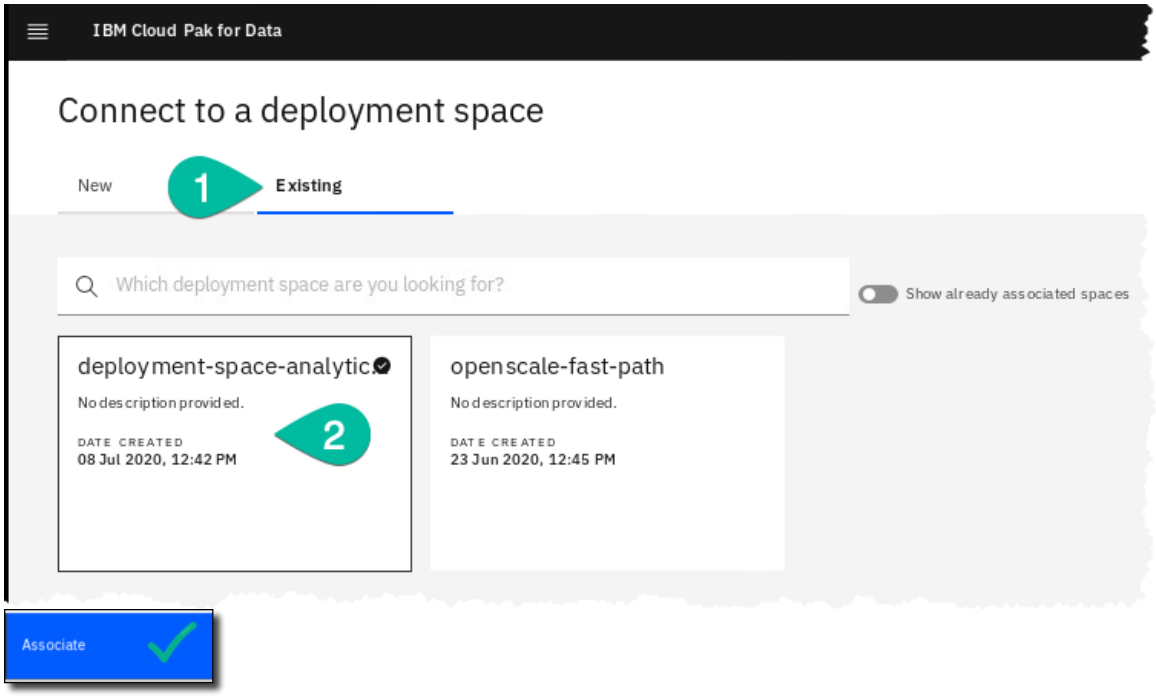
Notice it says this project is currently not associated with a deployment space.

Click Associate a deployment space,

Click section Existing,

Click Deployment Space deployment-space-analytics-project-workshop,

Click Associate,



Developer

Note: this deployment space was generated by running the notebook in the previous lab.

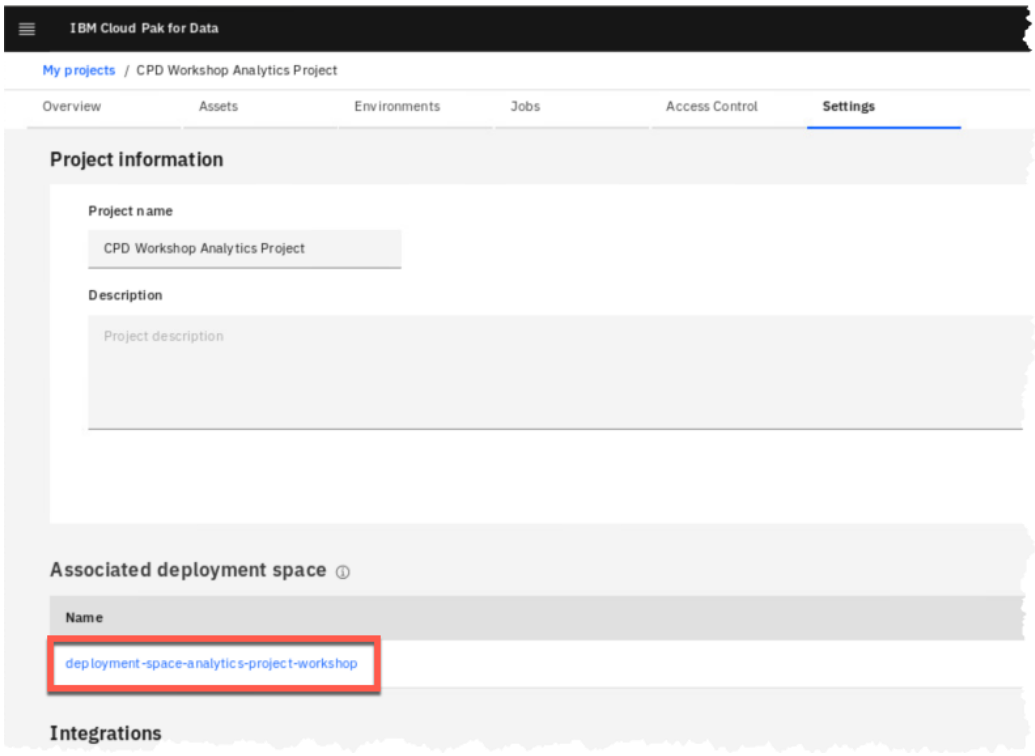
```
# Specify a names for the space being created, the saved model and the model deployment
space_name = 'deployment-space-analytics-project-workshop'

model_name = 'churn_risk_model'

deployment_name = 'churn_risk_model-deployment'
```

5.1 Create Deployment Space ¶

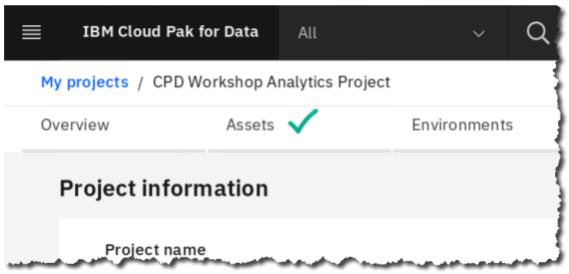
```
# create the space and set it as default
space_meta_data = {
        client.spaces.ConfigurationMetaNames.NAME : space_name
}
```

However, you could have created a deployment space through the CPD web client as well.

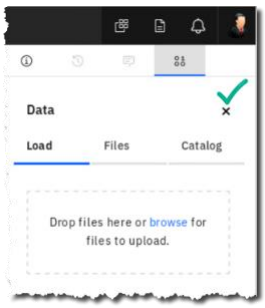
__8. The project should now show that it is associated with this deployment space. (You may need to scroll down to see this.)



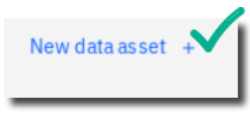
__9. Scroll up and click Assets,



__10. If the data asset add screen is open, close it.

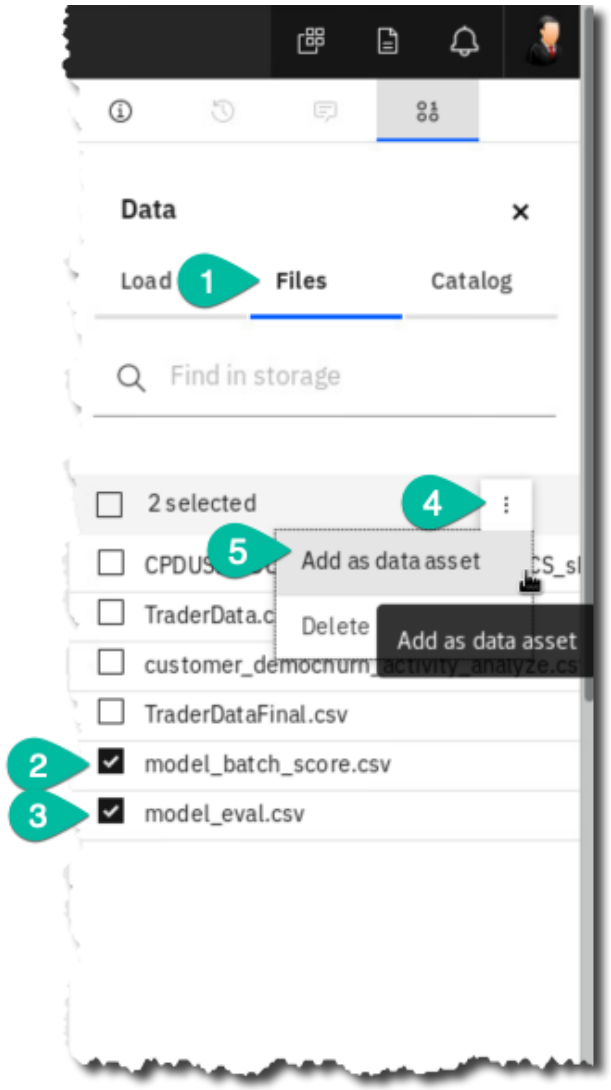


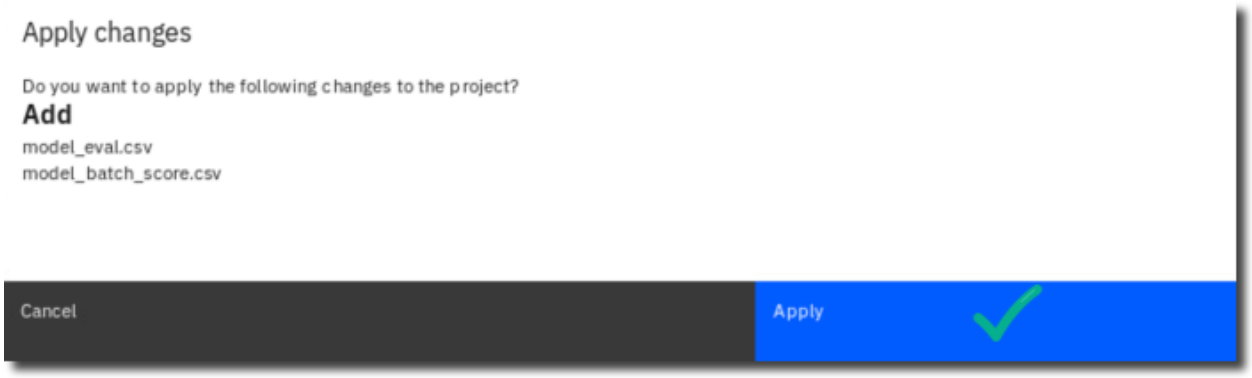
__11. Click + New data asset (This is how you open the data asset screen.)



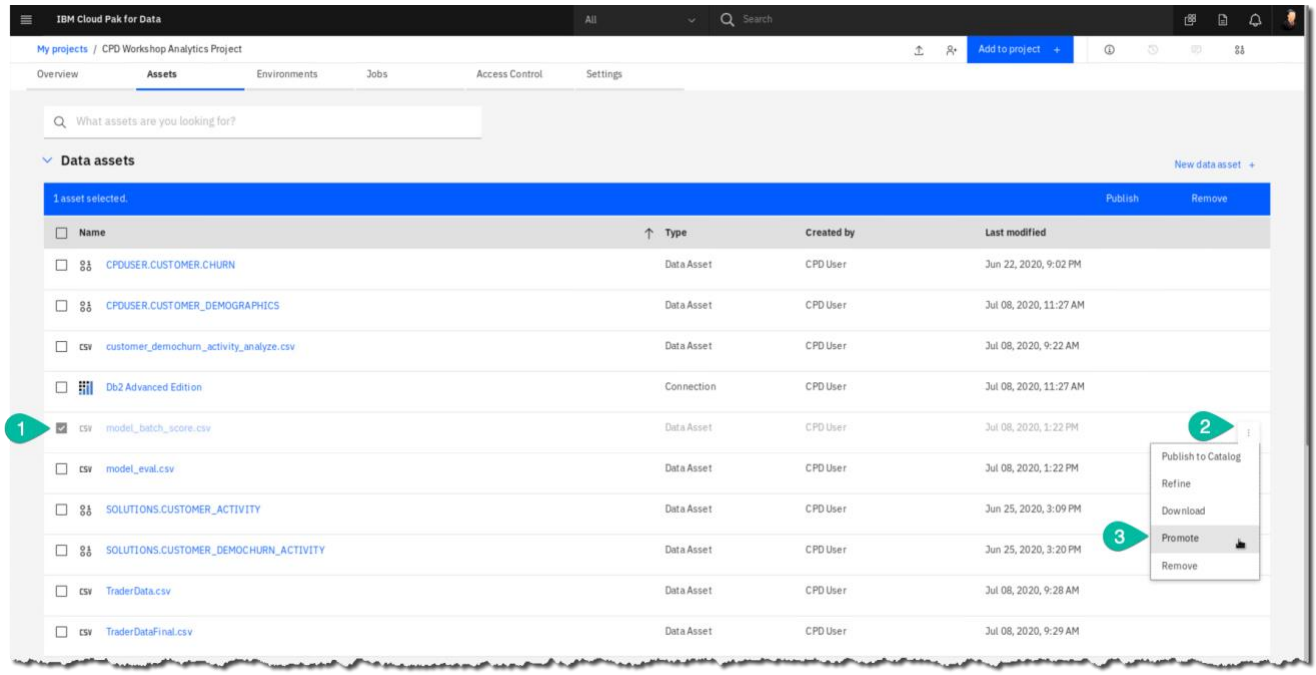
__12. Click Files, then select the two files: model_batch_score.csv and model_eval.csv.

(Note: these files were created by the notebook), then click ellipses ⋮ a Add as data asset and Apply.

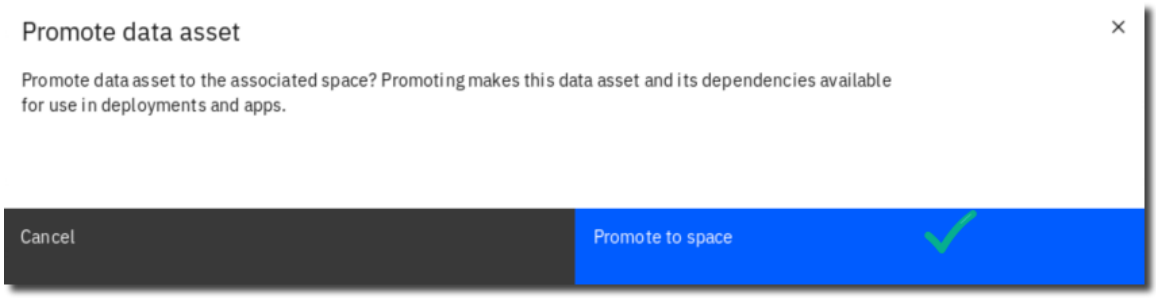




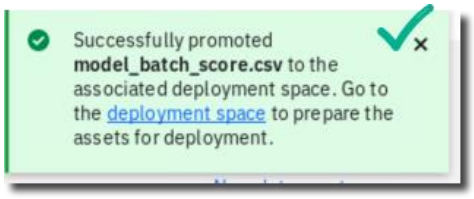
__13. Find the newly added data asset model_batch_score.csv, click on ellipses ⋮ and Promote.



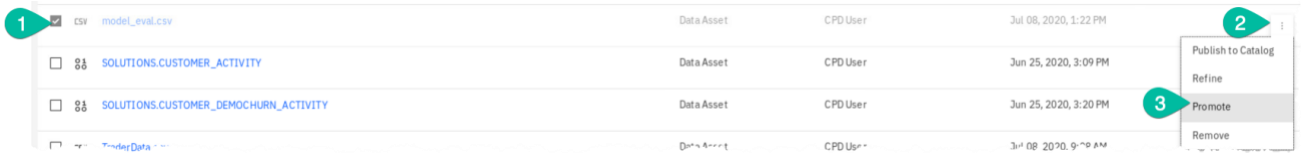
__14. Click Promote to space.



__15. The data asset is now promoted to the associated Deployment Space.



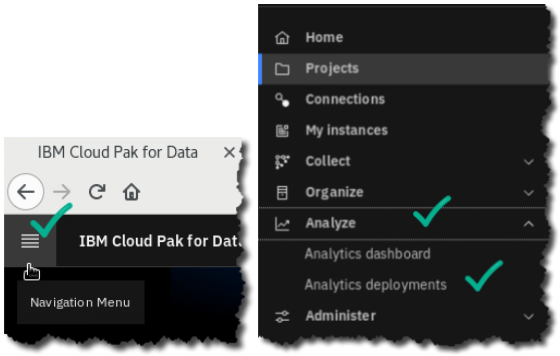
__16. Repeat for the data asset model_eval.csv.



7.4.2 Working in the deployment space

Now you can go to the deployment space to work in it:

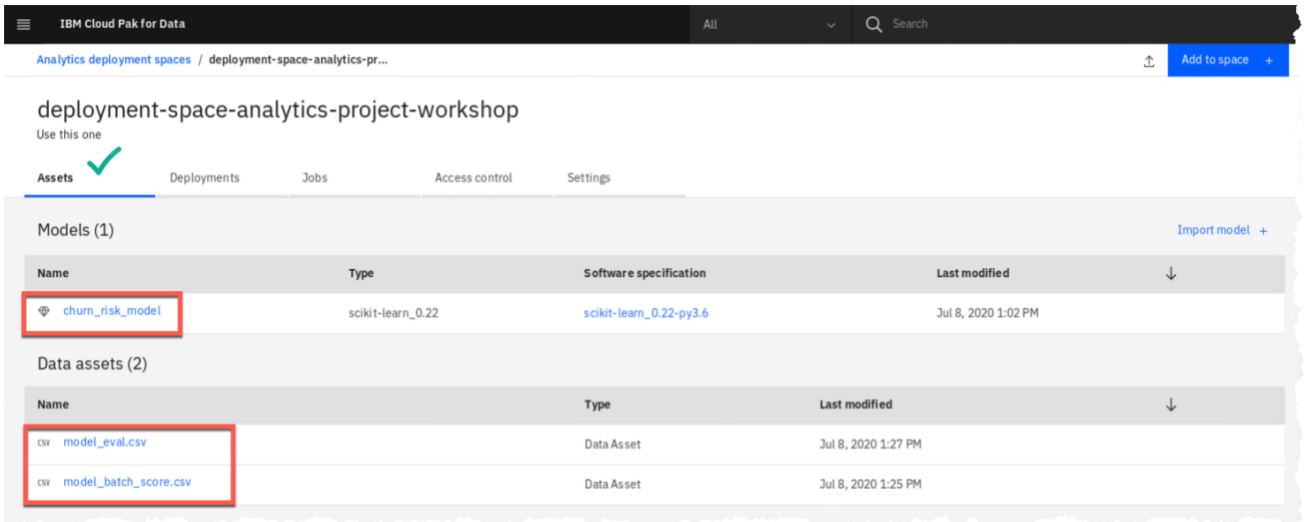
__17. Click Navigation menu ⇨ Analyze ⇨ Analytics deployments



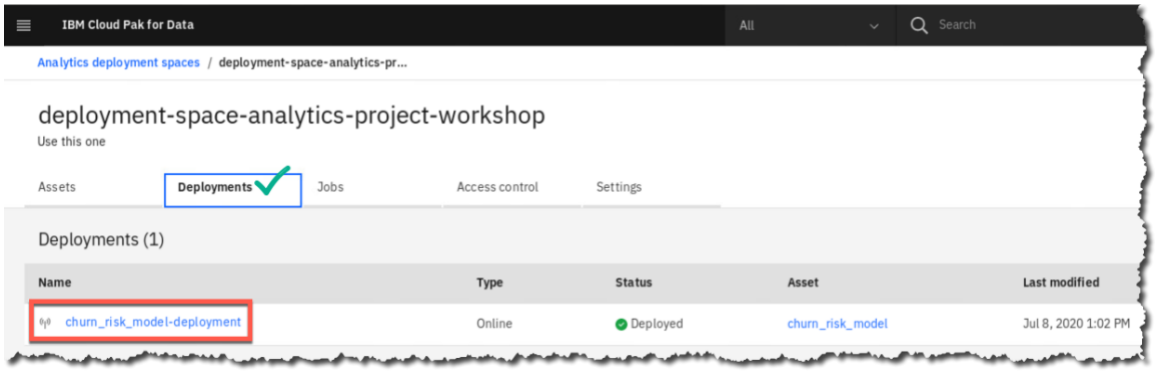
__18. In this screen, note that you could create a new deployment space if you wanted to. You will not need to, however, because the one we need was created by the notebook.

Instead, click deployment-space-analytics-workshop.

__19. In the Assets section, there is one Model: churn_risk_model (created by the notebook) and there are two Data Assets (the ones you just promoted).



__20. In the Deployments section, there is one deployment: churn_risk_model-deployment (which uses model churn_risk_model).



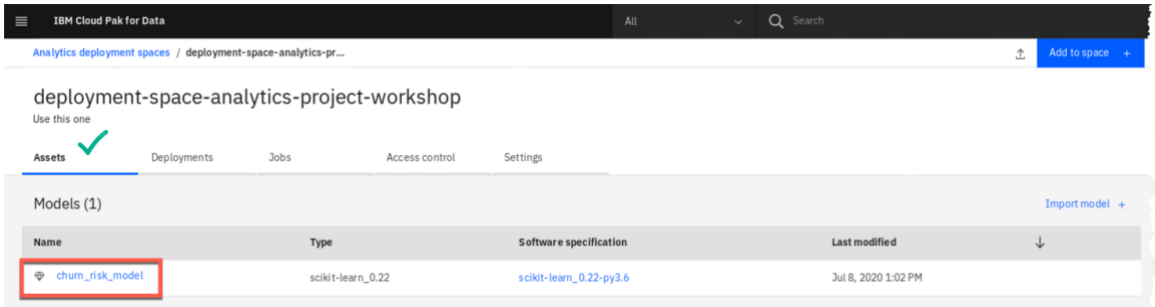
Developer

The notebook created all of these CPD assets – part of the code that did it is shown here.

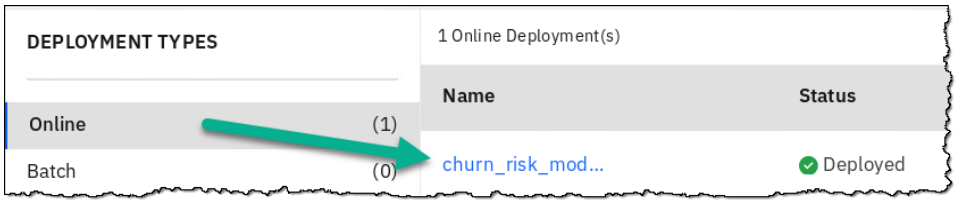
```
# Specify a names for the space being created, the saved model and the model deployment
space_name = 'deployment-space-analytics-project-workshop'

model_name = 'churn_risk_model'

deployment_name = 'churn_risk_model-deployment'
```

__21. Return to section Assets (click on it).

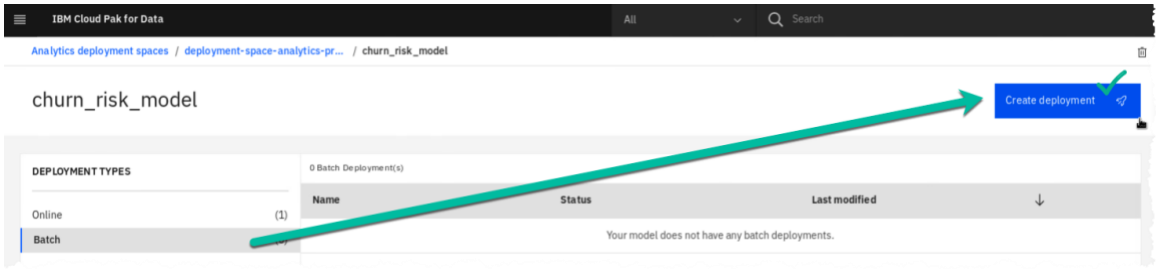
Click on the model: churn_risk_model.



__22. It shows one Online deployment (as seen previously in the Deployments section).

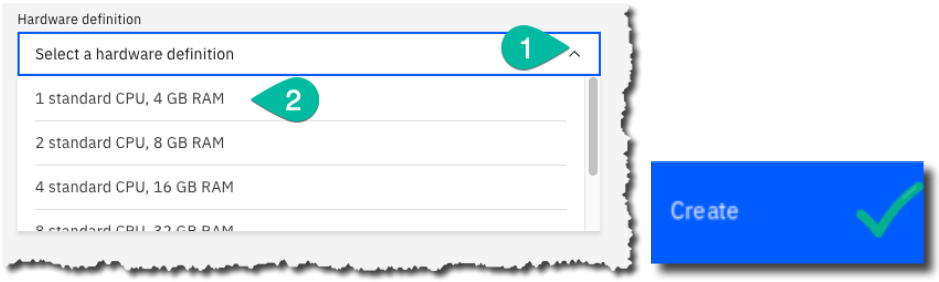


__23. Click Batch ⇨ Create Deployment

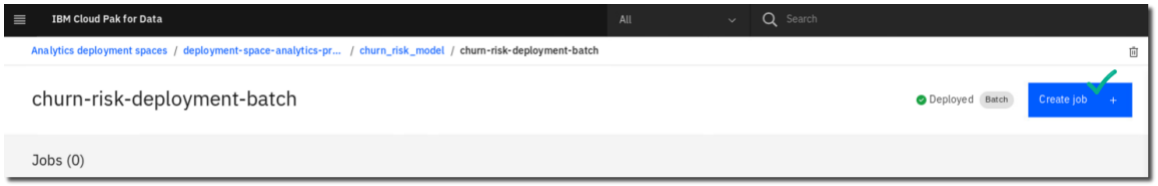


__24. Fill in name: churn-risk-model-deployment-batch.

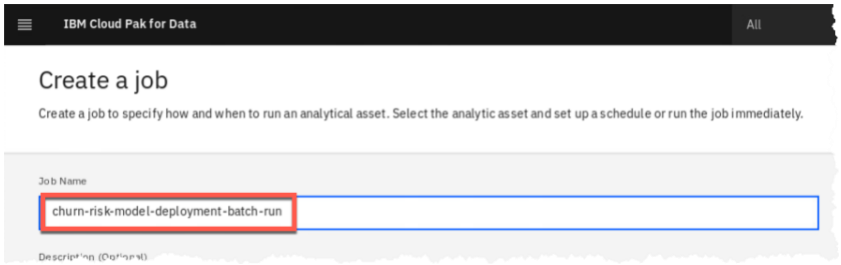
__25. Scroll down and find Select a Hardware definition. Select 1 standard CPU,4 GB RAM, and click Create.



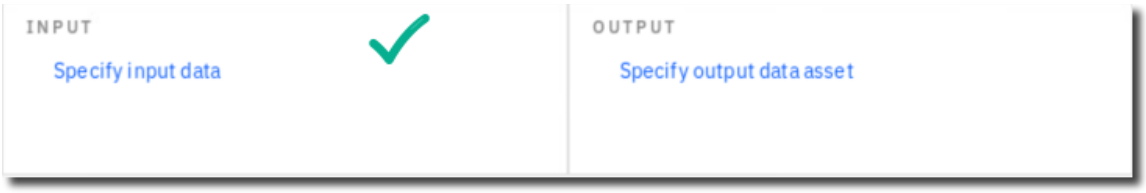
__26. You will be taken to the Jobs screen. Click Create job +.



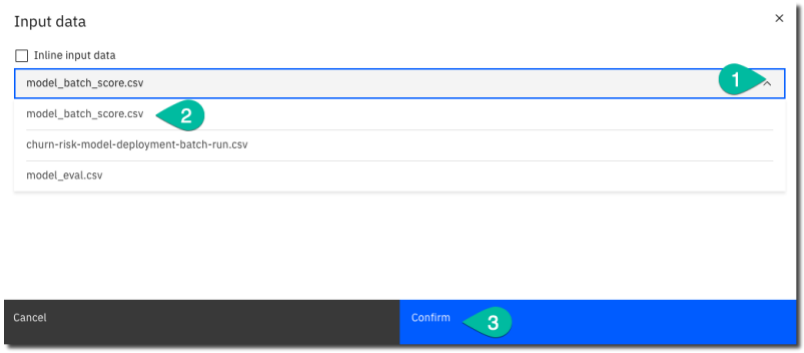
__27. Fill in the Job Name: churn-risk-model-deployment-batch-run.



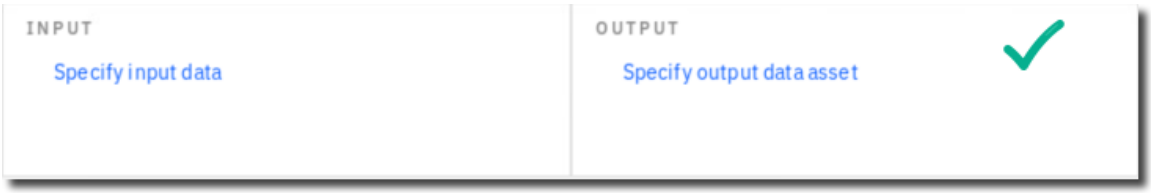
__28. Click on sections for INPUT – Specify input data.



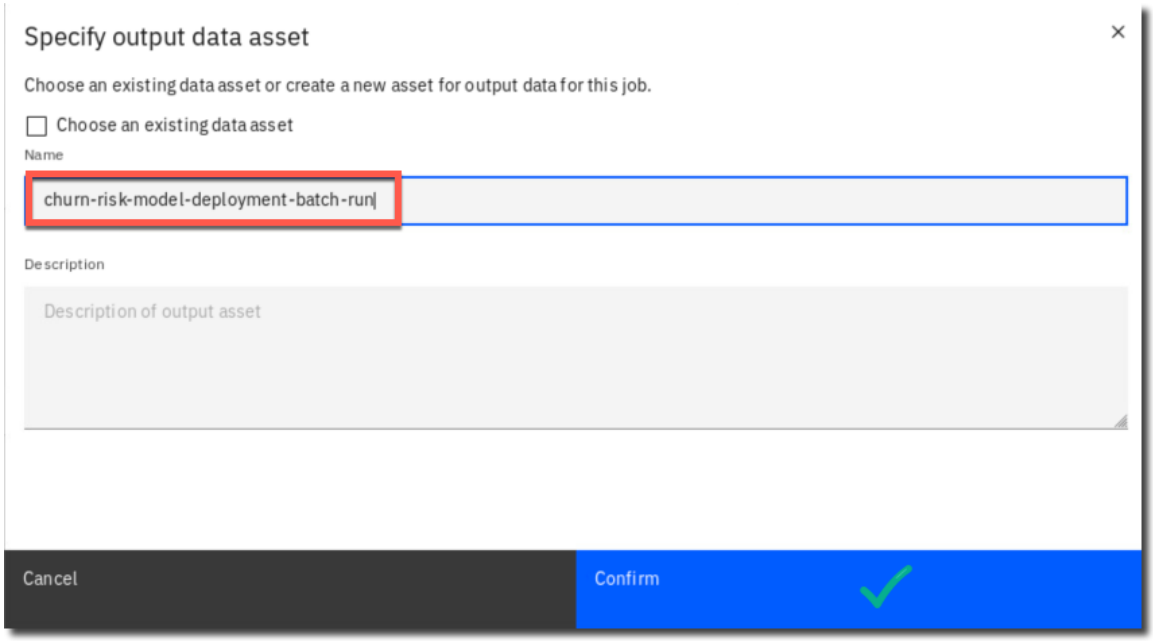
__29. Set the input value to model_batch_score.csv and Confirm.



__30. Select the OUTPUT – Specify output data asset.



__31. In this case, we will output to a new dataset. Enter churn-risk-model-deployment-batch-run for name and Confirm.

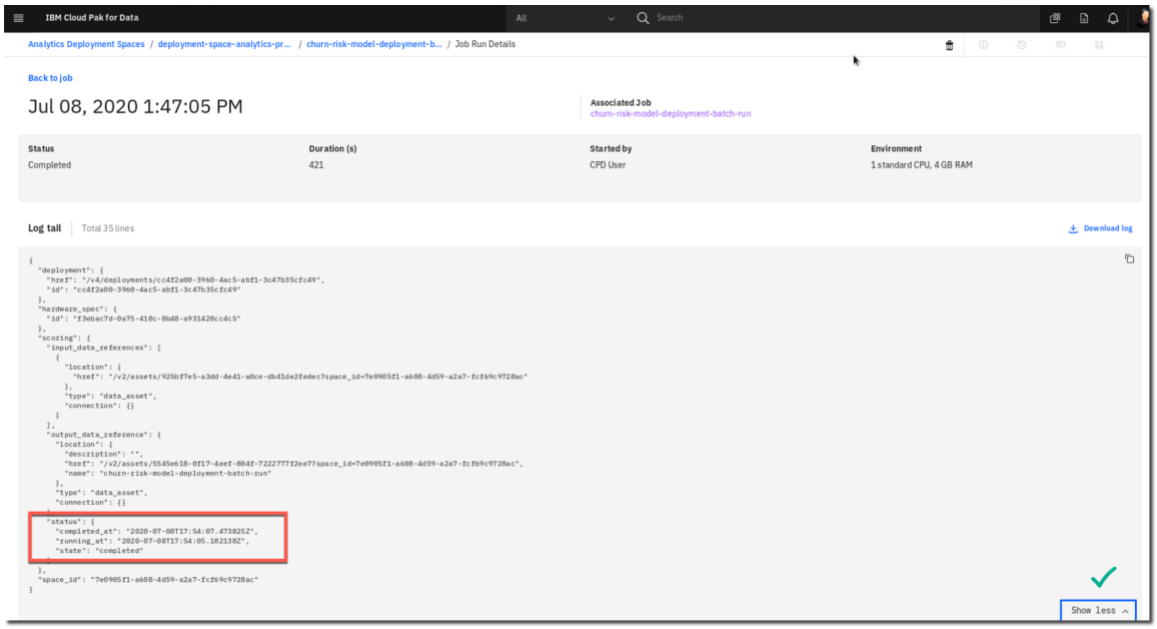


__32. Select Create and Run.

__33. The job will queue, run, and complete.
Note: You may need to refresh your browser to see the "Completed" result. You may have to wait around 5-10 minutes for the job to complete.

__34. Click on the Start Time value when Completed.

__35. Review the job – select Show More at bottom right.

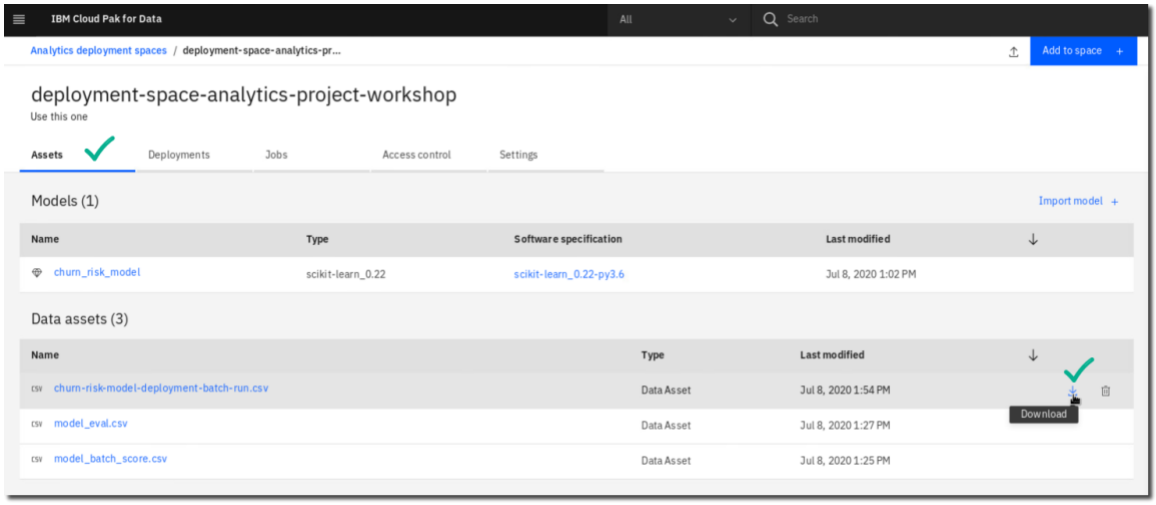


__36. After reviewing the job, click on the breadcrumb to return to the deployment space.

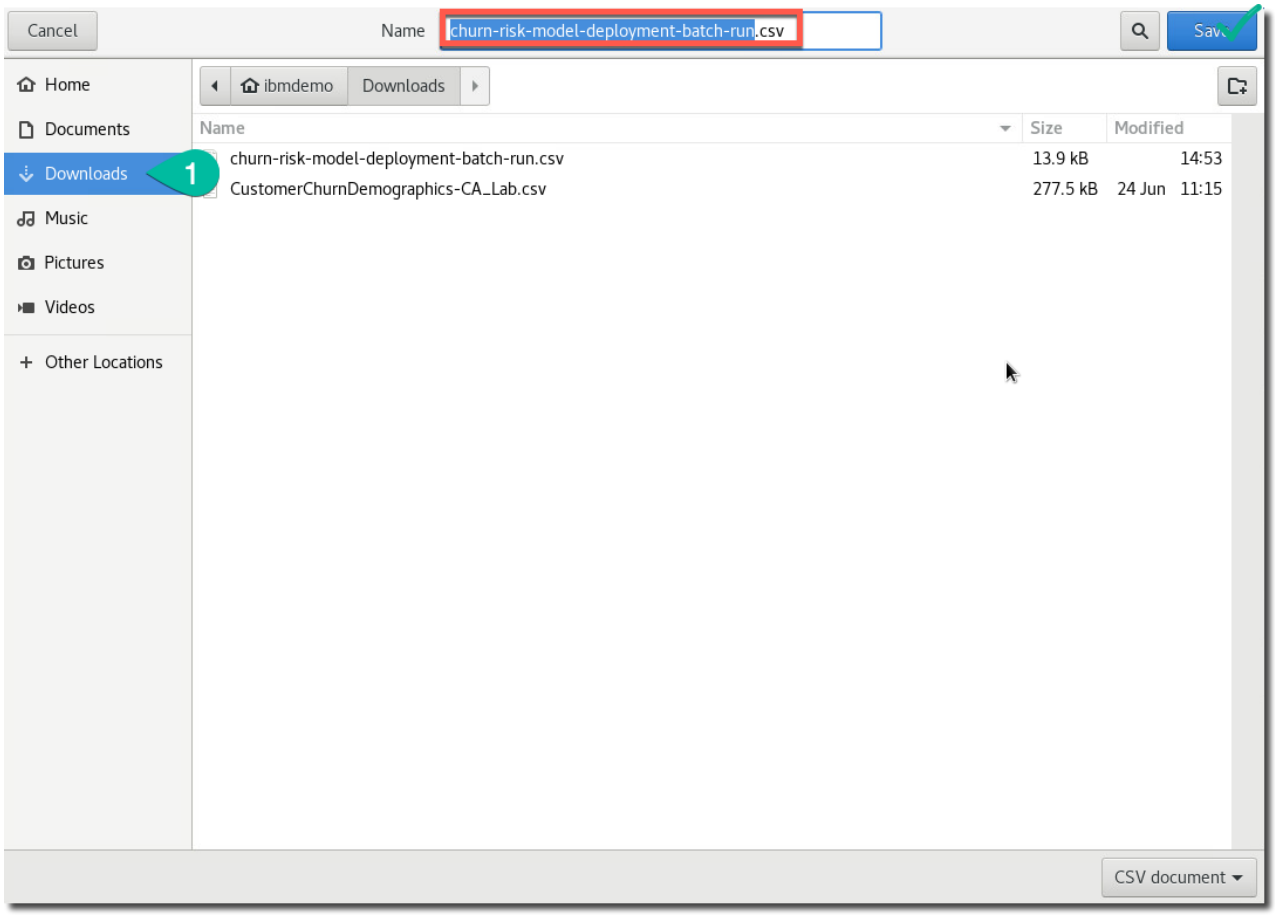
__37. In the section Assets, find the new output file under Data Assets.

The file name is churn-risk-model-deployment-batch-run.csv.

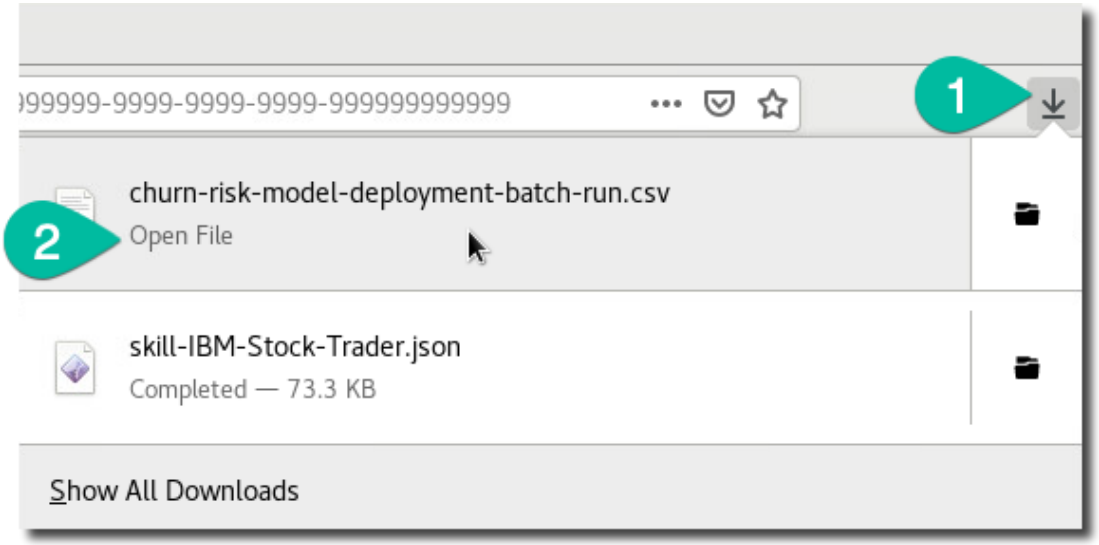
__38. On that file line, click Hover to uncover the download (down arrow) icon ⋮ and Click it.



__39. Save your File to the Documents folder and Save.

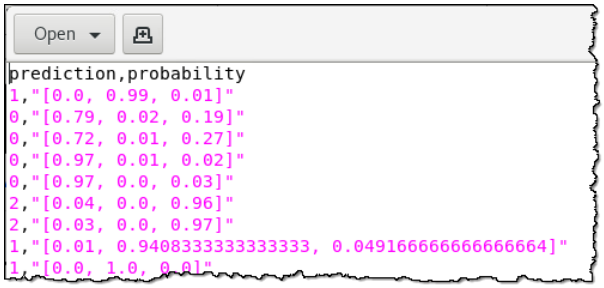


__40. Review the downloaded file by clicking on the download (down arrow) icon in your Firefox browser, then select the file.



*Note: this file is located on the operating system under **/home/ibmuser/Downloads.***

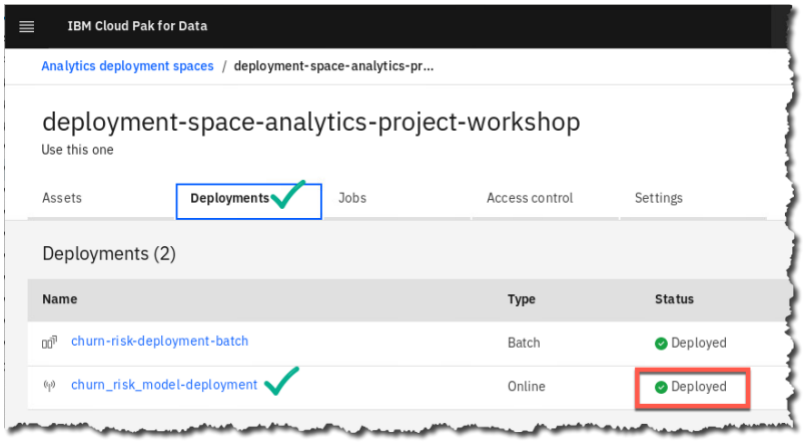
__41. The batch run output is available for review.



__42. Close the file edit windows after reviewing.

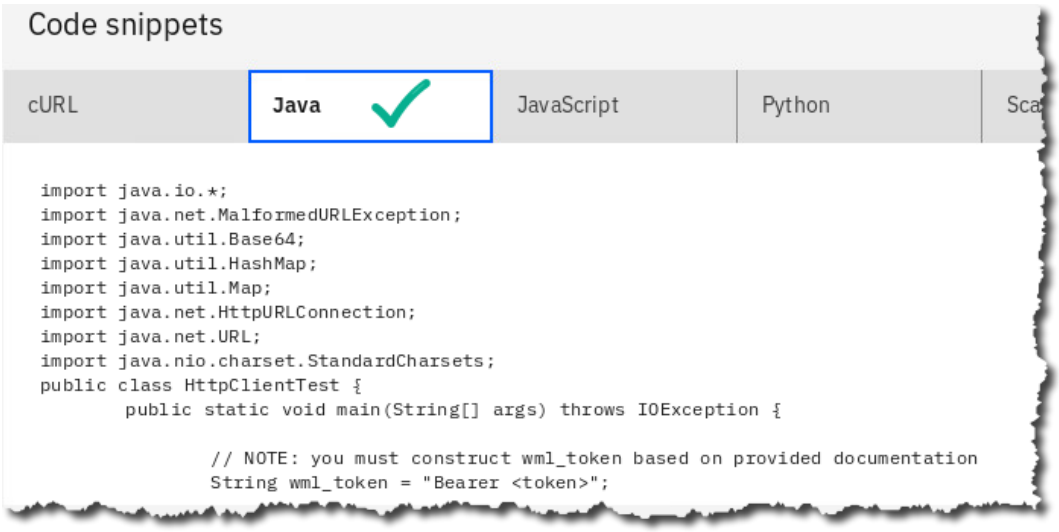
__43. Click Deployments.

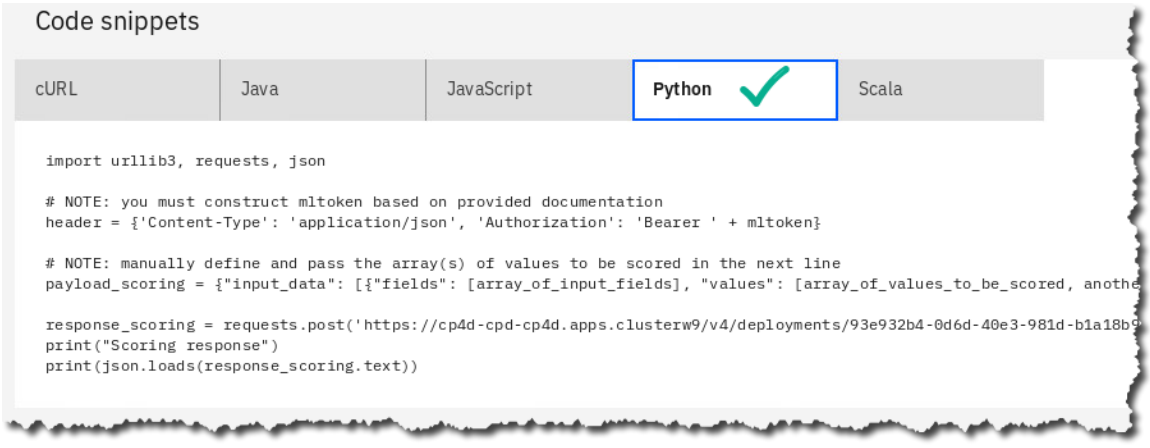
Click Online deployment: churn_risk_model_deployment.



__44. Under section API reference, review the Endpoint for this model. This is what your applications can use to call this model.

__45. Review the code snippets for each language that can aid a Developer to easily embed the model into an application.

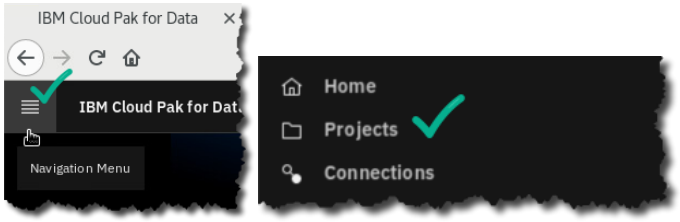




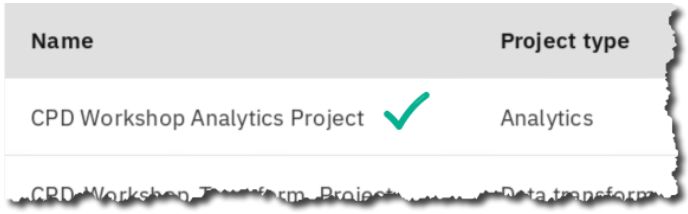
7.5 Deploying and testing the AutoAI model

7.5.1 Deploying the AutoAI model

__46. In the CPD web client, click the Navigation Menu ⇨ Projects.

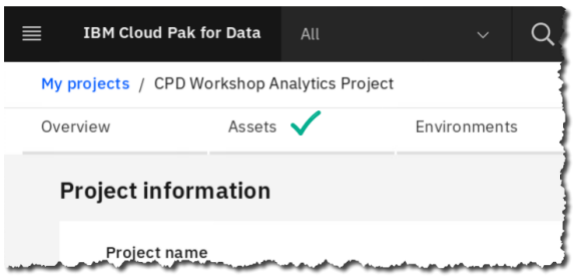


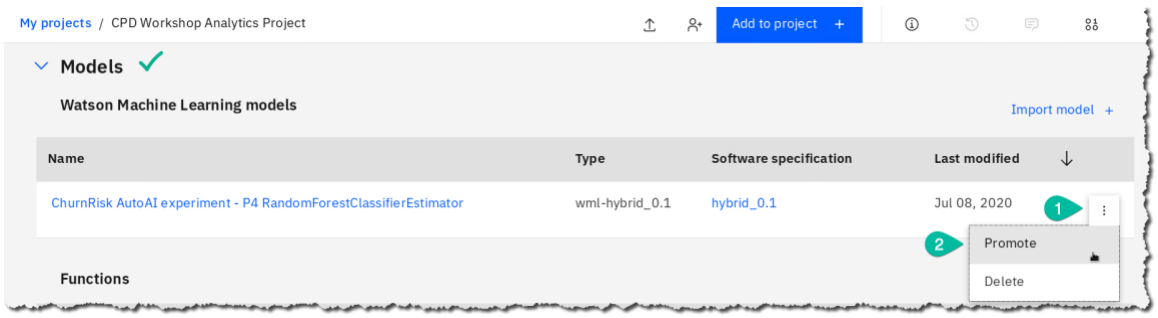
__47. Select the project: CPD Workshop Analytics Project.

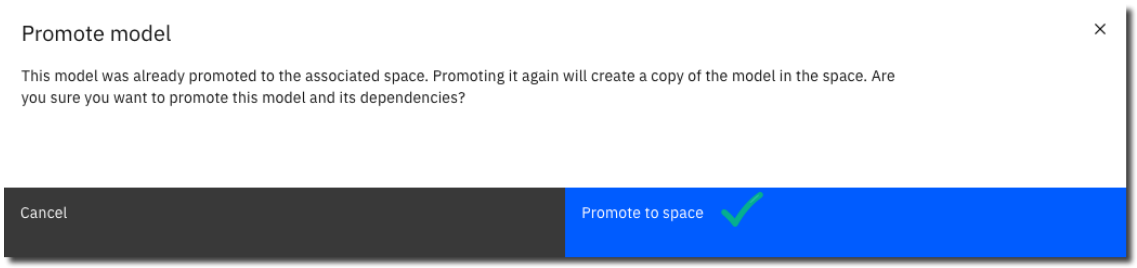


__48. Under Assets, scroll down to Models.

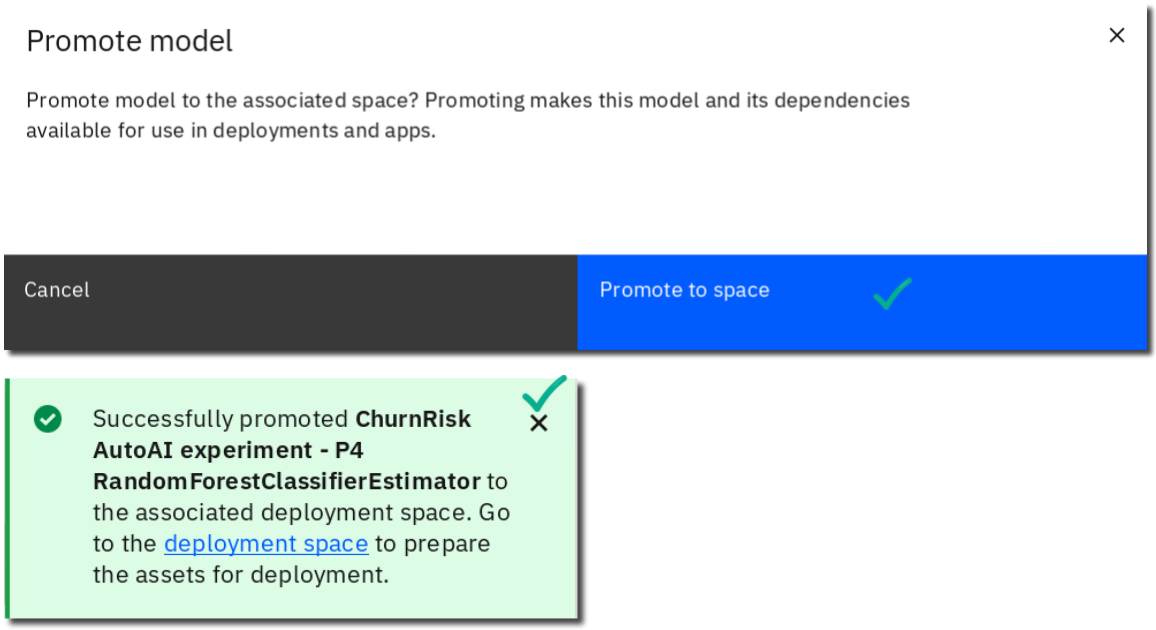
On the model created from the AutoAI experiment (ChurnRisk AutoAI experiment – P4 RandomForestClassifierEstimator), click on ellipses ⋮ then Promote and Promote to space.



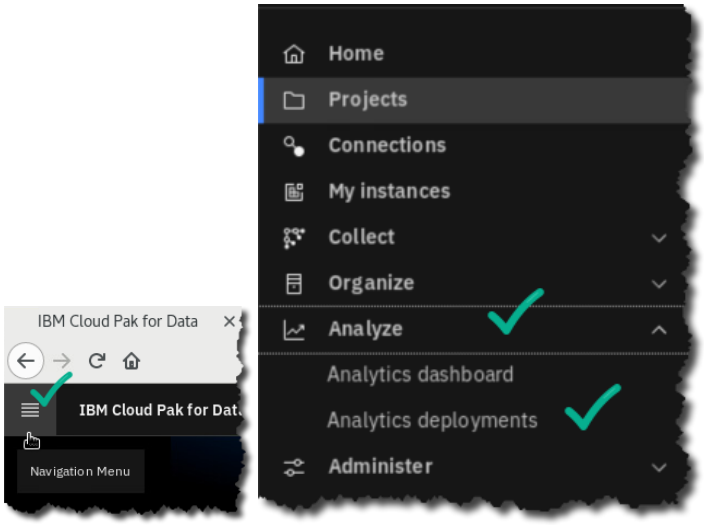




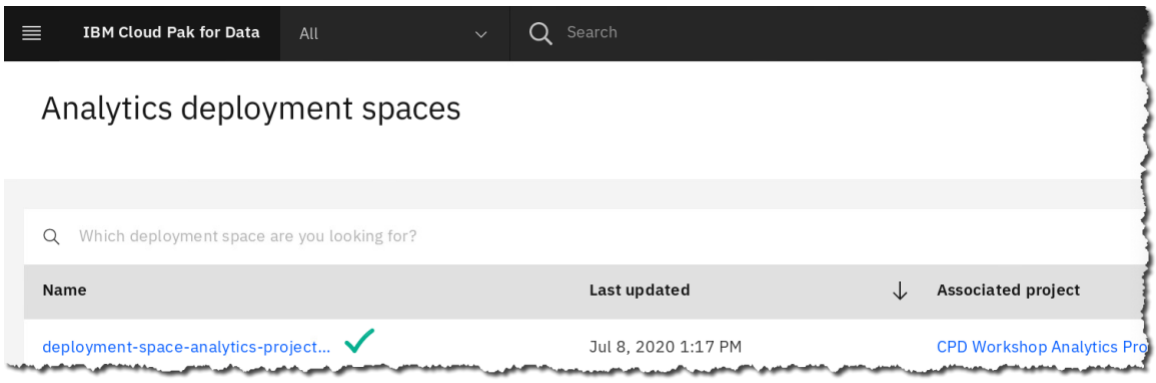
__49. This promotes the model to the associated Deployment Space



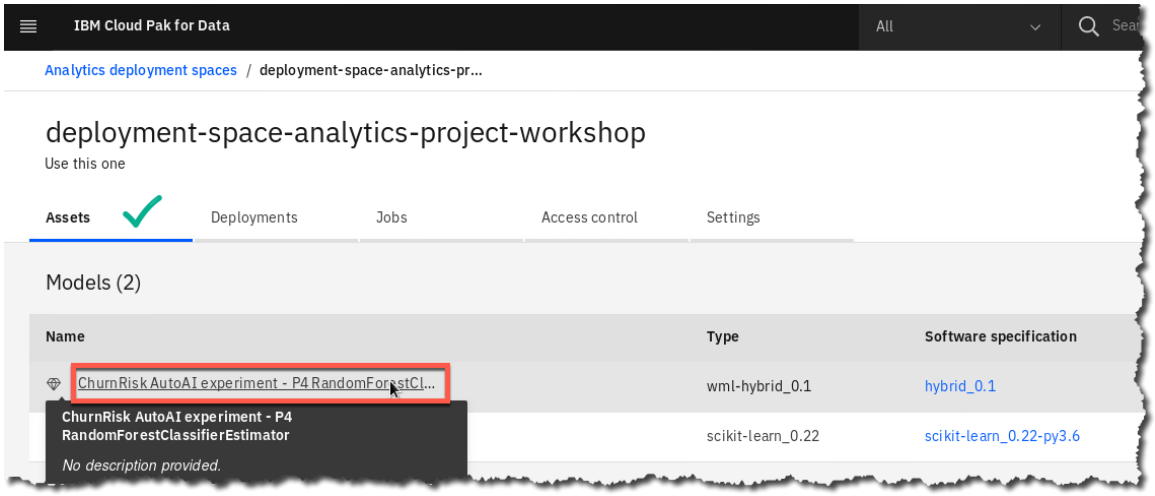
__50. Click Navigation menu ⇨ Analyze ⇨ Analytics deployments.



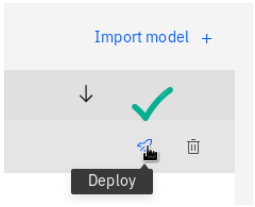
__51. Click deployment-space-analytics-project-workshop.



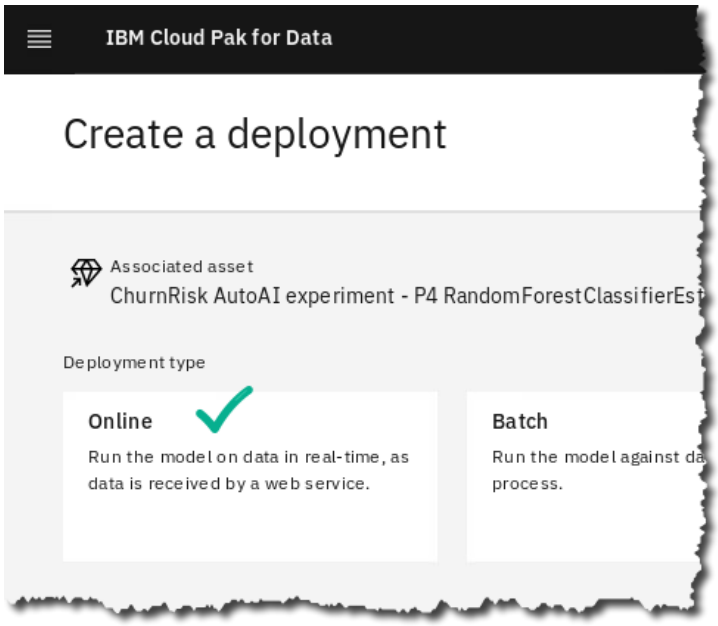
__52. Under section Assets, find the model you just promoted from your project. (*Don't click on it.)*



__53. At the end of the row for this model, click Hover and click on the Deploy (rocket) icon.

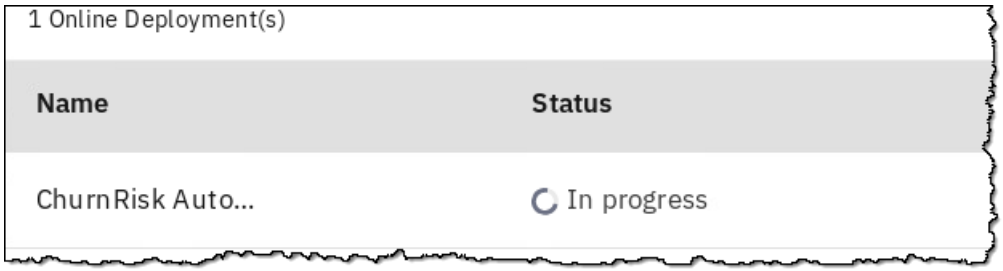


__54. Choose Online.

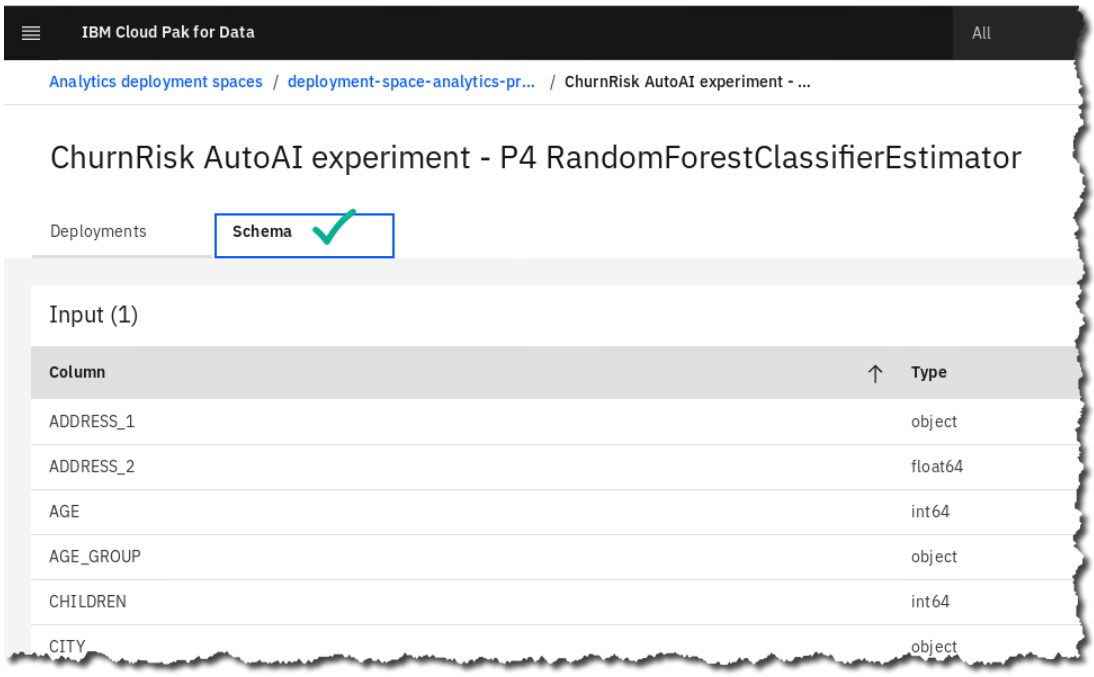


__55. Name: ChurnRisk AutoAI Experiment deployment a Create.

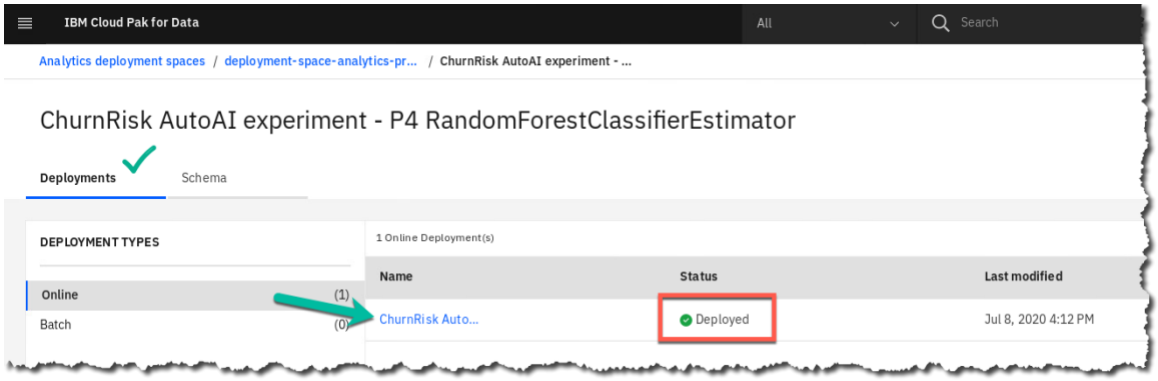
__56. You will see the online model being deployed.



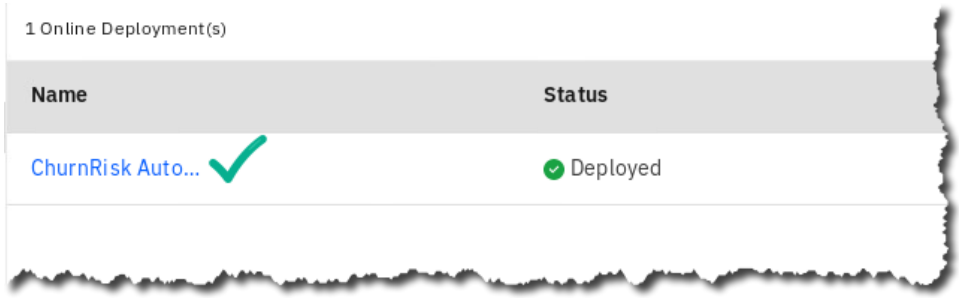
__57. While you are waiting for it to deploy, review the model Schema.



__58. Select Deployments. The Online model is now deployed in short time with zero coding.



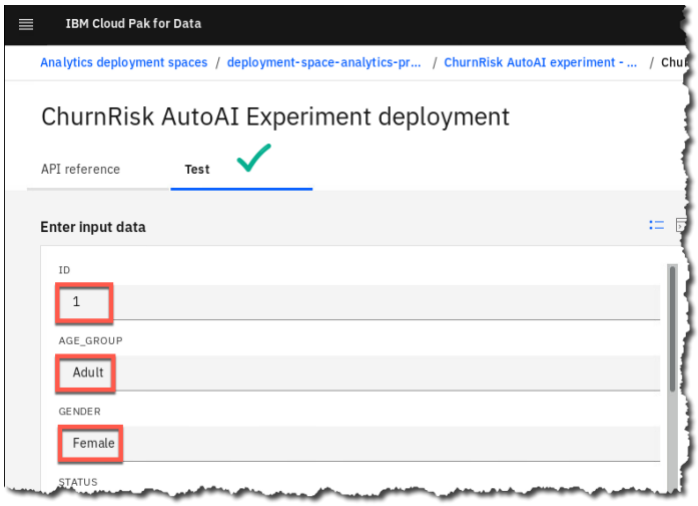
__59. Click on the deployment.



7.5.2 Testing the AutoAI model

__60. Click on tab Test.

Fill in the first three input data fields as shown below.



__61. Scroll down and click on Predict.

(Note: filling in more fields would make the prediction more accurate, but we only did three fields to show you the process.)

__62. Your machine learning prediction and probability is returned.

7.6 Lab conclusion

Today, organizations need better, faster and more integrated methods for model deployment that can include online, batch and virtual testing. The Cloud Pak for Data deploy capabilities help serve this need so they can more easily infuse AI into their applications.

In our scenario, the Trade Co. developer used the platform to easily deploy and test models.

**** End of Lab 07 - Deploy**

Lab by Burt Vialpando and Kent Rubin, IBM

Lab 08 Infuse: Watson OpenScale

8.1 Lab overview

Business value is achieved when a model is infused into an application, regularly monitored and updated. But most organizations need help with that. They want to know when the accuracy of a model begins to decline so it can automatically be corrected. They also want to know how their models make decisions, and make sure they eliminate bias.

The first part of this lab will leave the Trade Co. use case to focus on a Credit Risk use case demo that is built into Watson OpenScale. It shows how OpenScale can detect and mitigate bias while providing explainable output for the model's decision.

In this `Infuse` lab you will learn how Watson OpenScale assists you in this.

8.2 Persona represented in this lab

The `Developer` persona is the likely role to perform the various `Infuse` tasks in this lab. However, the Data Scientist persona could perform these tasks as well.

Persona (Role)	Capabilities
 Developer	Developers create and maintain the end-user applications that utilize the output from all the other personas on the CPD platform.

8.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Double-click the desktop icon: `Cloud Pak for Data Web Client.`

__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign in`.

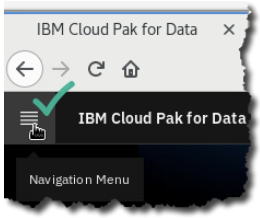
8.4 Credit Risk built-in demo

8.4.1 Run OpenScale Auto setup

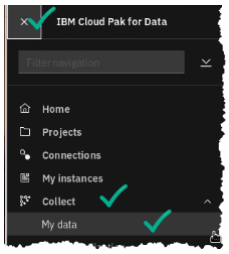
Watson OpenScale provides a quick setup utility that will automatically set up a data mart, create a model, and record seven days' worth of measurements into the OpenScale monitors. This creates the Credit Risk demo that we will go through in this lab.

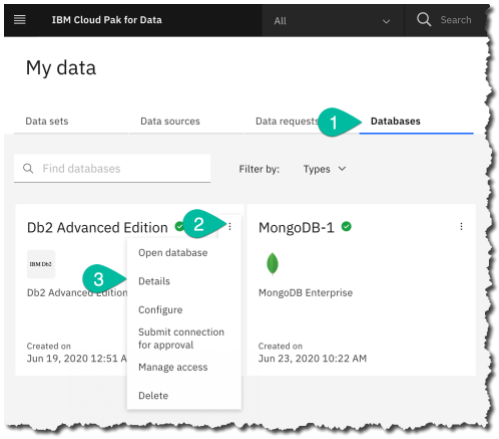
__4. Start by first retrieving the Db2 Advanced Edition connection information.

Click on the `Navigation Menu` ("hamburger" icon) at the top left of the screen.



__5. Click `Collect` ⇨ `My Data` ⇨ `Databases` ⇨ `Db2 Advanced Edition` ⇨ `ellipsis` ⋮ ⇨ `Details.`



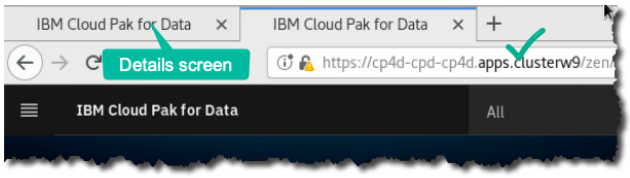


__6. Scroll to find the Access Information. Notice Username, Password, Host, Port and Database.

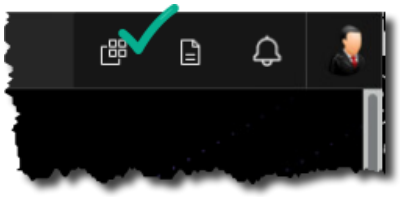
__7. If you are using the remote desktop, click on the icon for `Cloud Pak for Data Web Client;` otherwise, duplicate your existing browser tab.

or


__8. There should now be two web client browser tabs open. The first tab has the `Details Screen` for the Db2 Advanced Edition just reviewed. The second is the new web client browser page just launched.

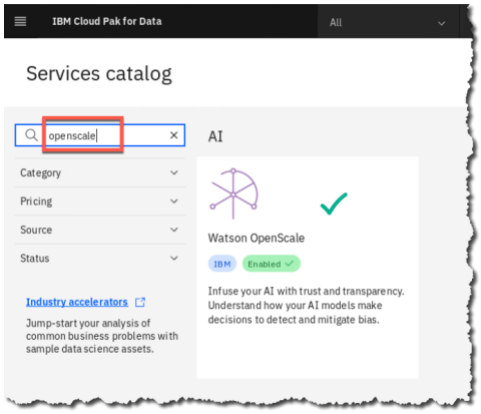


__9. From the second (new) web client browser tab, click the icon for `Services` (at the top right corner of the screen)

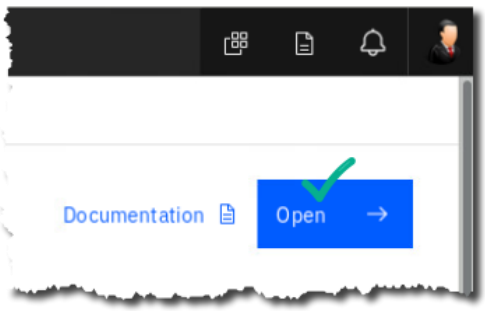


__10. In the search window, type `openscale.`

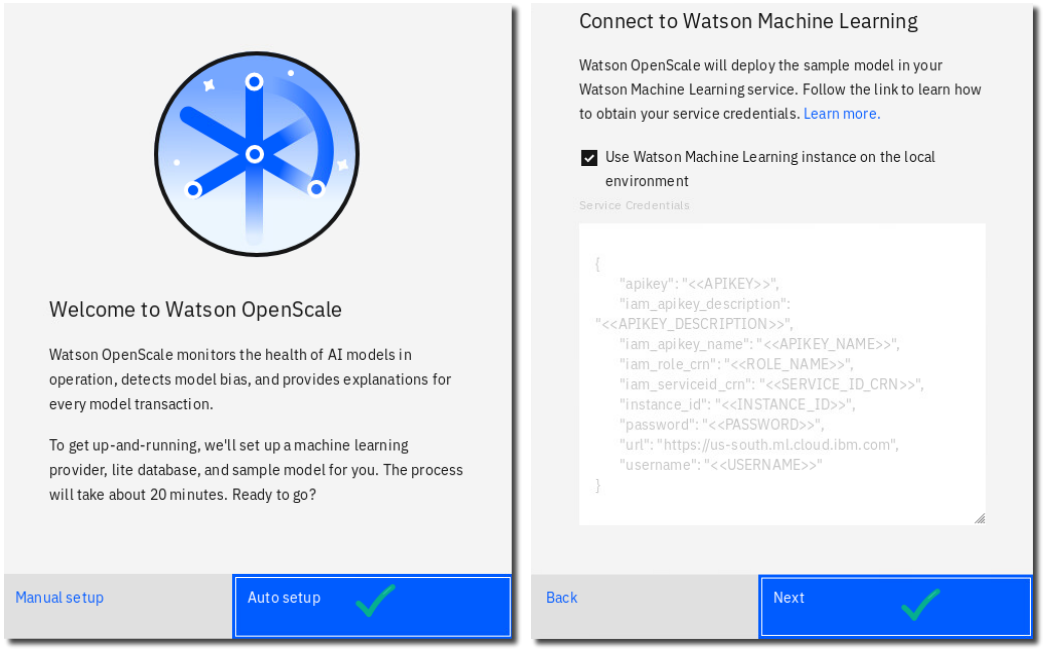
Click on tile `Watson OpenScale.`



__11. Click `Open.`



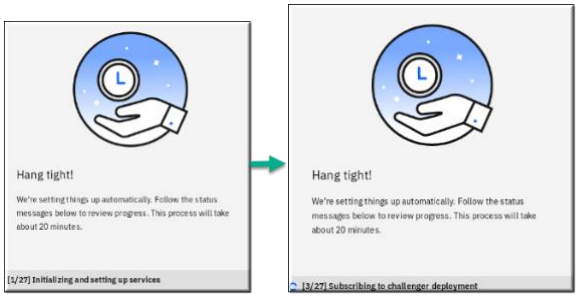
__12. Click `Auto setup` ⇨ `Next.`



__13. Fill in the connection information here from the `Db2 Advanced Edition Details` screen you have opened in the other tab. (*Note: Your hostname may vary.*)

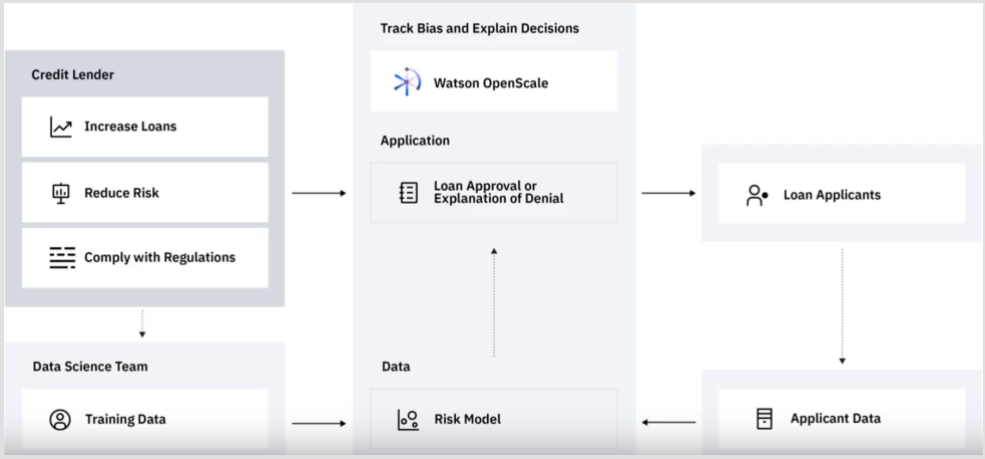
When all filed in, click `Prepare.`

__14. The Auto setup will take 20 minutes or so to run. During this time, read about the OpenScale monitors and the Credit Risk scenario it is creating as it goes through the 27-step setup.



Developer

For this demo, you will be monitoring a model that attempts to predict credit risk based on demographic data as well as credit history, residence information, age, employment status, and more. The scenario and model use synthetic data based on the [UCI German Credit dataset.]



Developer

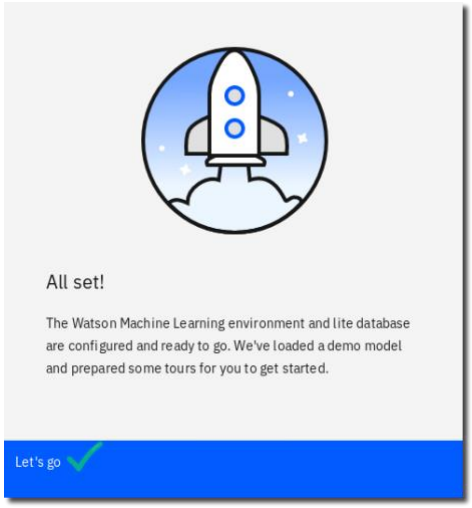
Credit lenders can monitor risk models for performance, bias and explainability to limit risk exposure from regulations and create more fair and explainable outcomes for customers.

Traditional lenders are under pressure to expand their digital portfolio of financial services to a larger and more diverse audience, which requires a new approach to credit risk modeling. Their data science teams currently rely on standard modeling techniques - like decision trees and logistic regression - which work well for moderate datasets and make recommendations that can be easily explained. This satisfies regulatory requirements that credit lending decisions must be transparent and explainable.

To provide credit access to a wider and riskier population, applicant credit histories must expand beyond traditional credit, like mortgages and car loans, to alternate credit sources like utility and mobile phone plan payment histories, plus education and job titles. These new data sources offer promise, but also introduce risk by increasing the likelihood of unexpected correlations which introduce bias based on an applicant's age, gender, or other personal traits.

The data science techniques most suited to these diverse datasets, such as gradient boosted trees and neural networks, can generate highly accurate risk models, but at a cost. Such "black box" models generate opaque predictions that must somehow become transparent, to ensure regulatory approval such as Article 22 of the General Data Protection Regulation (GDPR), or the federal Fair Credit Reporting Act (FCRA) managed by the Consumer Financial Protection Bureau.

__15. When the Auto setup has completed, click `Let's go`. You need not go through the Guided Tour, but you can if you have the time.



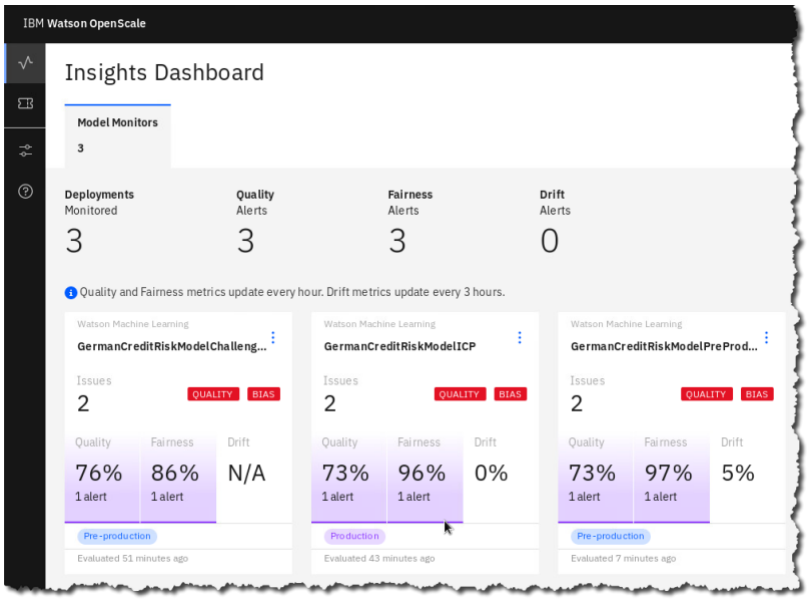
8.4.2 Explore the fairness and quality monitors

Watson OpenScale provides two types of monitors: application monitors, and model monitors.

We will begin with the model monitors. The model monitors section of the Insights Dashboard provides an overview of all the models being monitored by OpenScale.

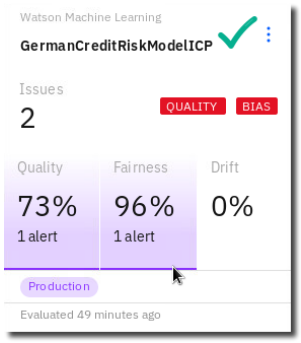
You can monitor models hosted with Watson Machine Learning on public or private clouds, Microsoft Azure, Amazon SageMaker, or custom models that provide JSON prediction output.

__16. The dashboard shows how many deployments are currently being monitored, as well as an overview of the alerts from those models. Below, each model is represented by a tile showing the machine learning provider and alerts for that model. The left side of the screen shows all of the active monitors for the models, divided into sections for `Fairness`, `Quality`, and `Drift`.

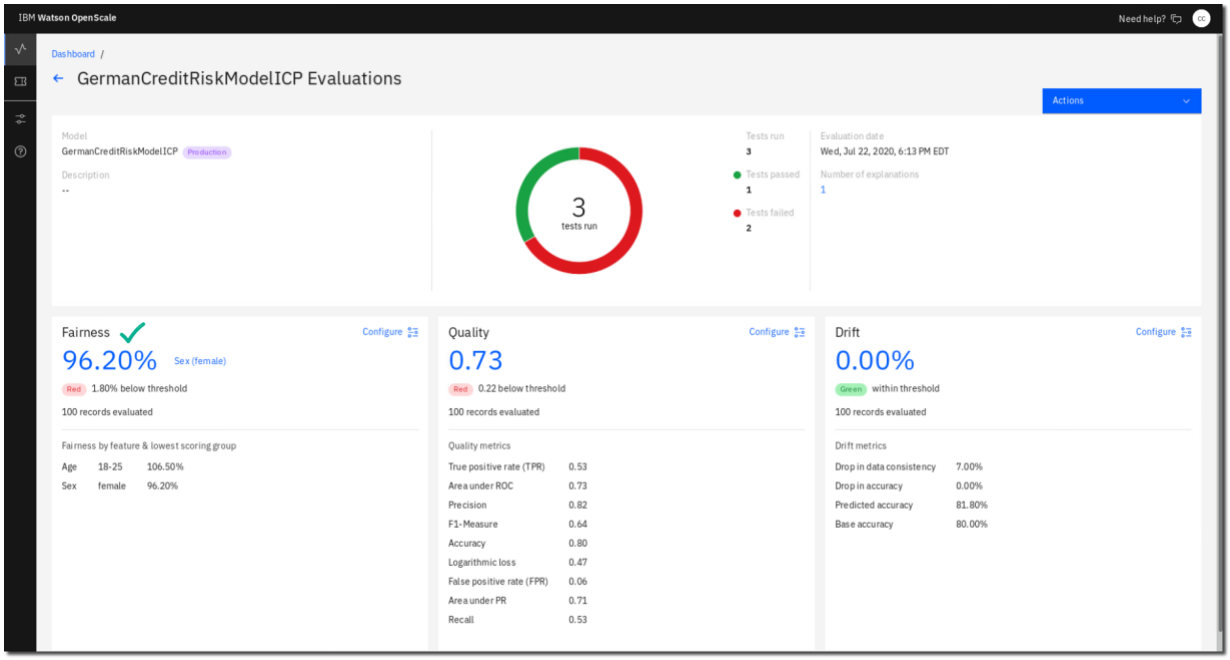


NOTE: Your actual values may vary slightly from what is shown in these lab screen shots as the data source has some randomness built into it.

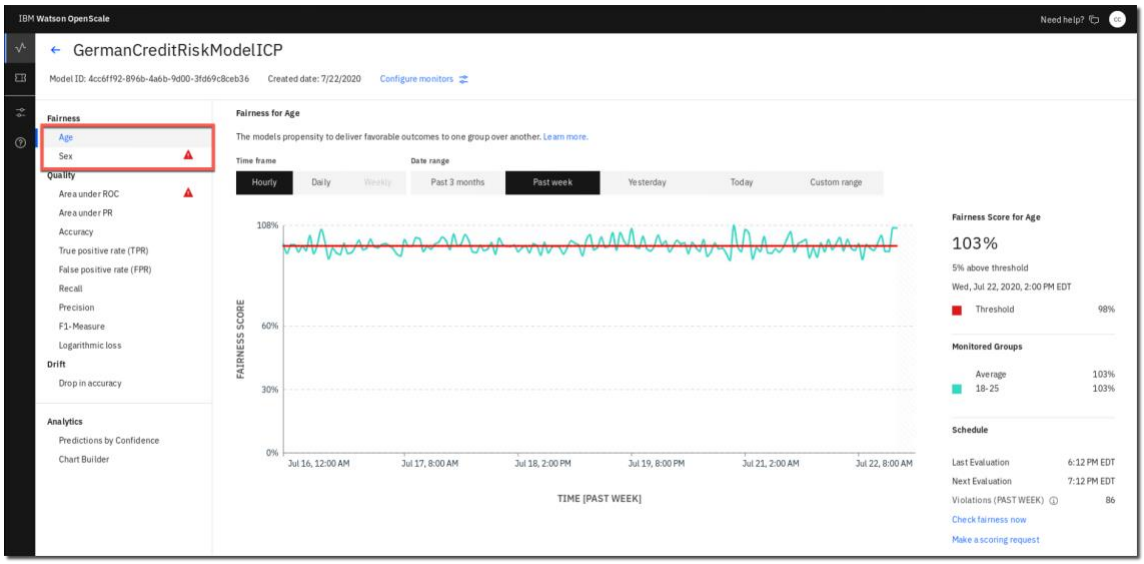
__17. Click on the tile `GermanCreditRiskModelICP.`



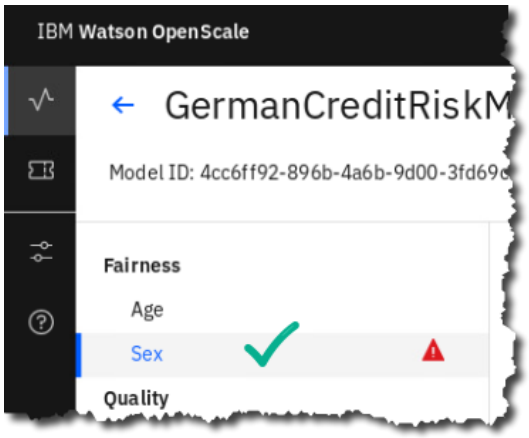
__18. You are presented with a dashboard of your model evaluations. Select the `Fairness percentage value link` in the Fairness tile.



__19. In the Fairness section, you can see that we have chosen to model two features, Age and Sex, for fairness. Additionally, you can see that we have an alert for the Sex feature.

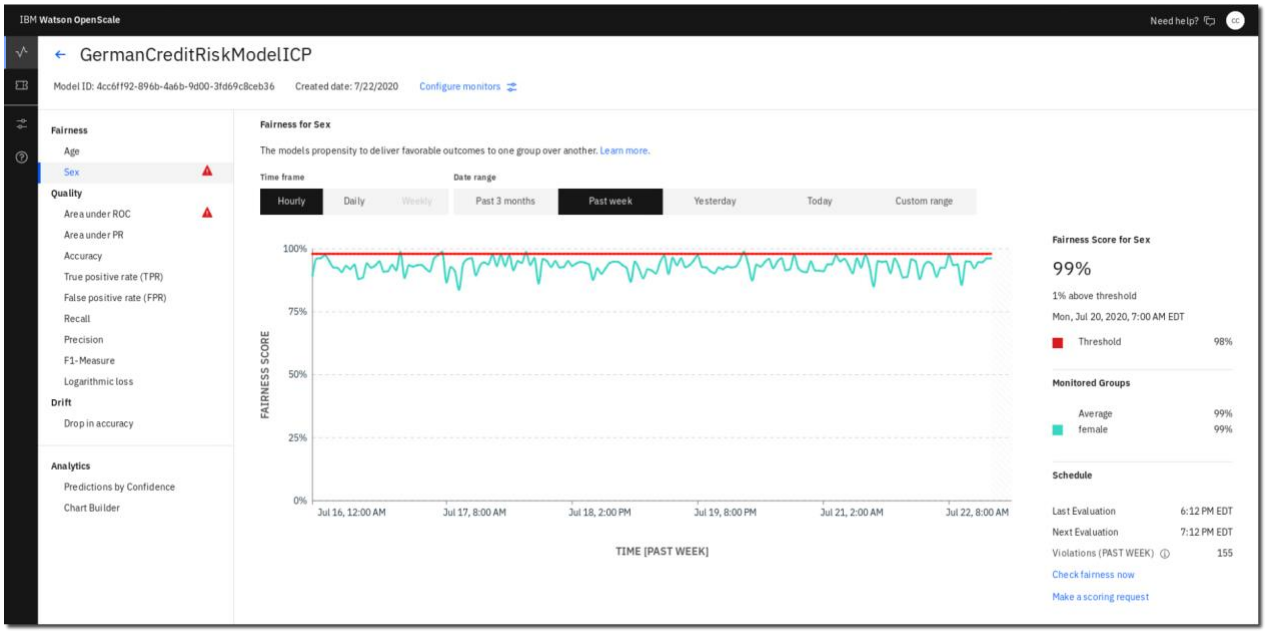


__20. Click on the fairness monitor `Sex.`

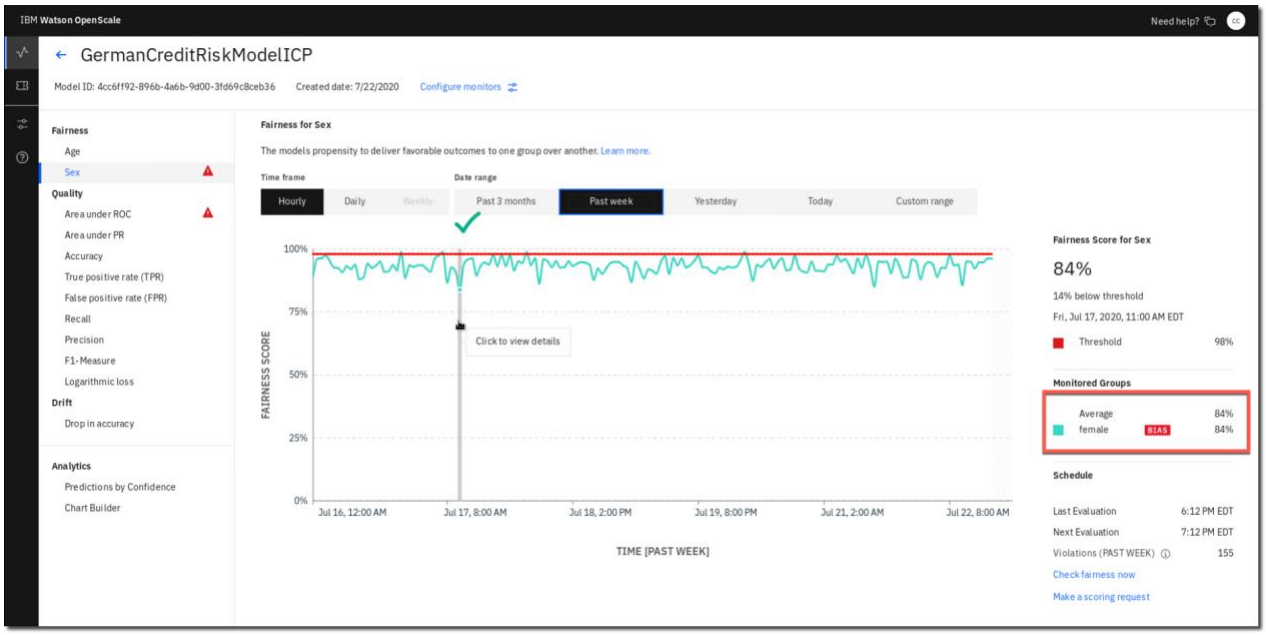


__21. Note the time frame above the graph. We are looking at hourly data from the past week, but the time frame can be adjusted as necessary. The graph shows the fairness score for females as a green line. The threshold we have set for an alert is shown by the red line.

As you can see, the fairness score has dropped beneath the threshold consistently over the past week, alerting us to a potential unfair bias issue with the model. As you move your cursor inside that date range, values are updated to show the Fairness Score for that date and time.



__22. Move your mouse on the chart to the point where fairness score was at its lowest (84%)



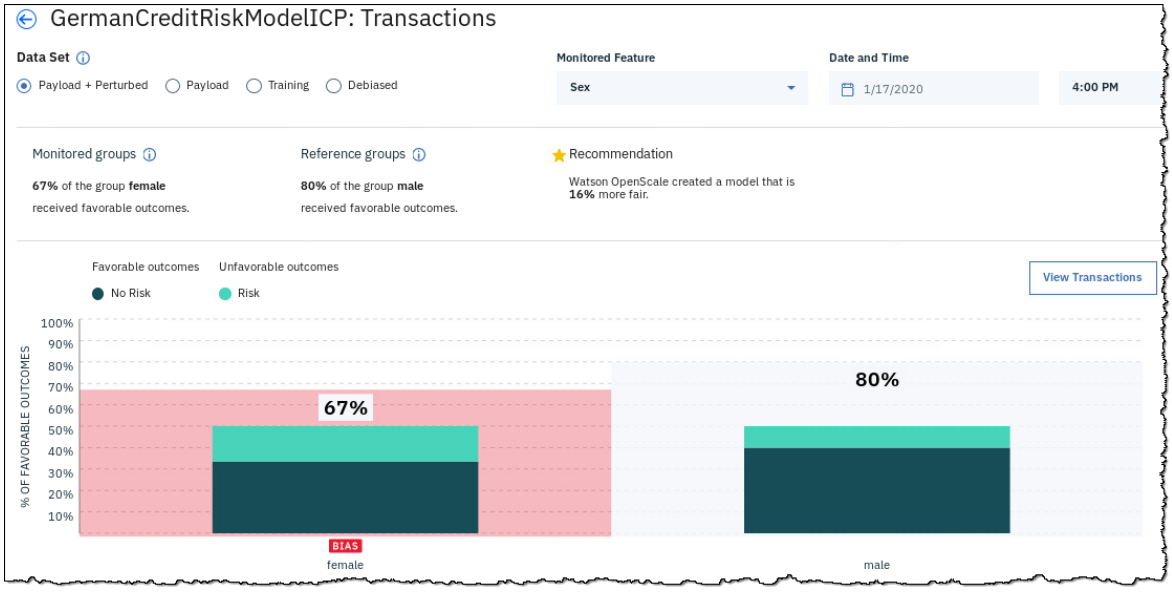
This screen shows us fairness details for this particular time period. OpenScale calculates its fairness score using a combination of actual predictions (payload data) and perturbed data,

generated when the prediction probability is close to 50%. OpenScale will flip the monitored feature to see how it affects the prediction outcome.

__23. Click on that lowest score line when it says: `Click to view details.`

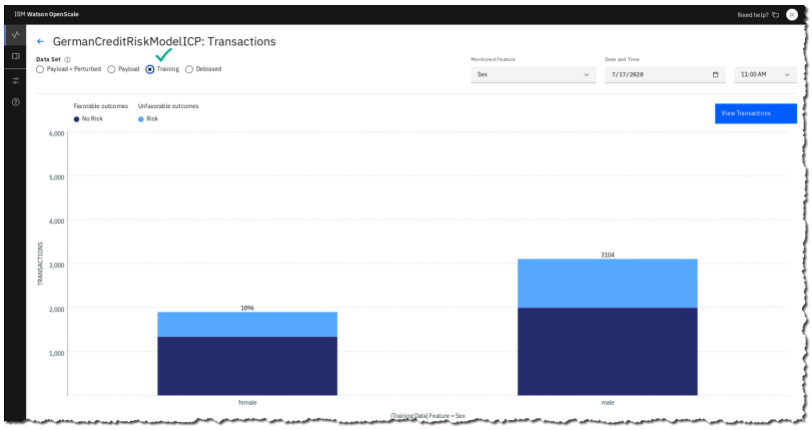
__24. The fairness score is reached by dividing the percentage of positive outcomes for the monitored group (females, 67%) by the percentage of positive outcomes for the reference group (males, 80%).

The graph shows the breakdown of positive and negative predictions for our two groups. You can use the radio button at the top to view payload and perturbed data, actual payload (prediction) data, and training data.



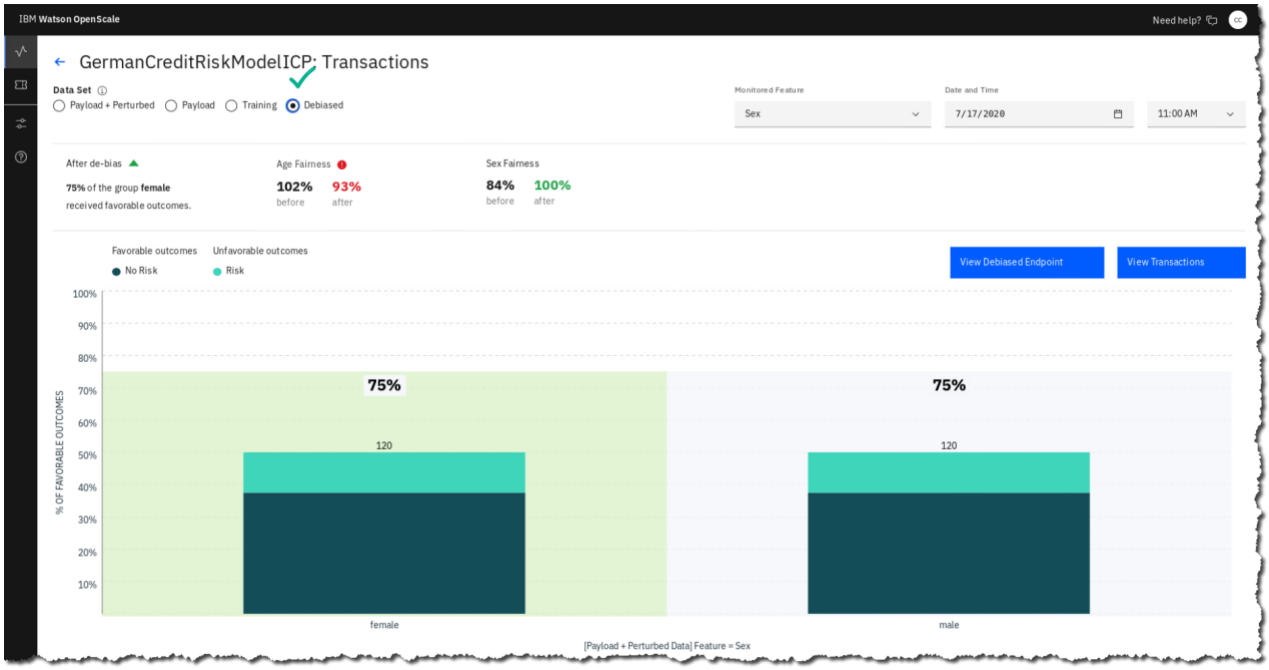
__25. Click data set (radio button) `Training` to view the training data breakdown.

As you can see, our training data had significantly more records for males than for females, which may be a potential source of our unfair bias against females.

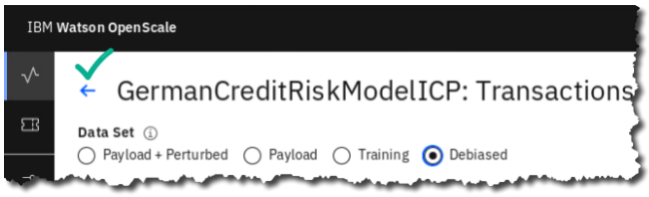


__26. Click the data set (button) `Debiased.`

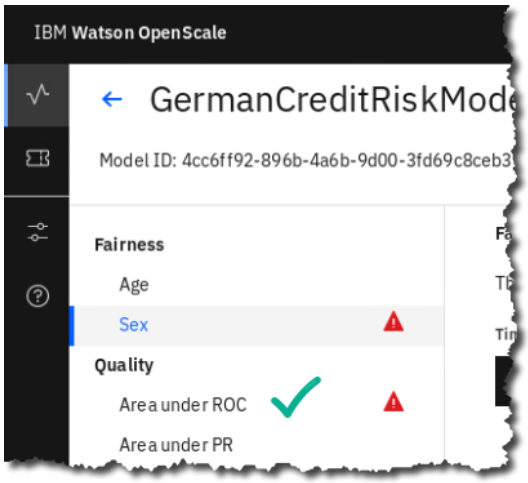
OpenScale can create a sort of “corrective lens” to reduce or remove unfair model bias. It does this by training another model to predict when an outcome of the production model is likely to be unfairly biased, flipping the feature value from the monitored group (female) to the reference group (male) and returning this prediction. On this screen, you can see how using this model will affect the fairness scores for other features.



__27. Click the `back arrow` to return to the model dashboard.



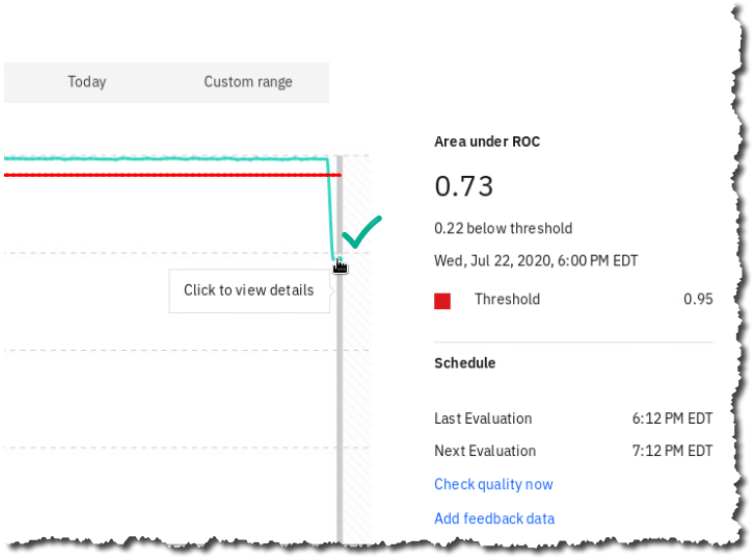
__28. Click the monitor `Area under ROC` in the section `Quality.`



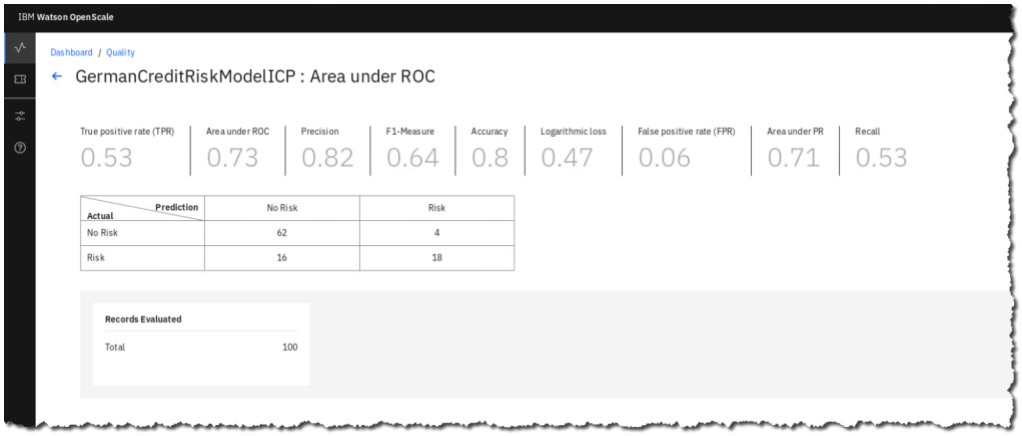
__29. OpenScale provides several different quality measurements. For our binary classification model, *Area under ROC* provides the best standard for model quality. These scores are generated by providing ground truth feedback data to the model, either via CSV upload or using a RESTful endpoint provided by OpenScale.

As with the fairness monitor, the chart in the middle of the screen shows model performance over an adjustable time window, with the relevant measurement shown as the green line and the alert threshold represented by the red line. As you can see, our model quality has consistently been above the threshold until the most recent measurement, represented by the right-most portion of the chart.

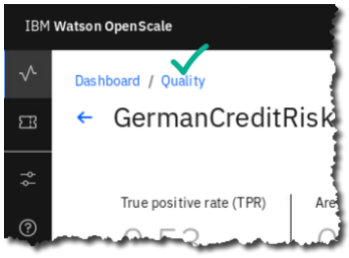
__30. Click on the right-most portion of the chart, where Area under ROC drops to `0.73`.



__31. Here we can see a further breakdown of the feedback data and the various accuracy scores, and the number of feedback records evaluated.



__32. Click on the `Quality` link to return to the GermanCreditRiskModelICP `Dashboard.`

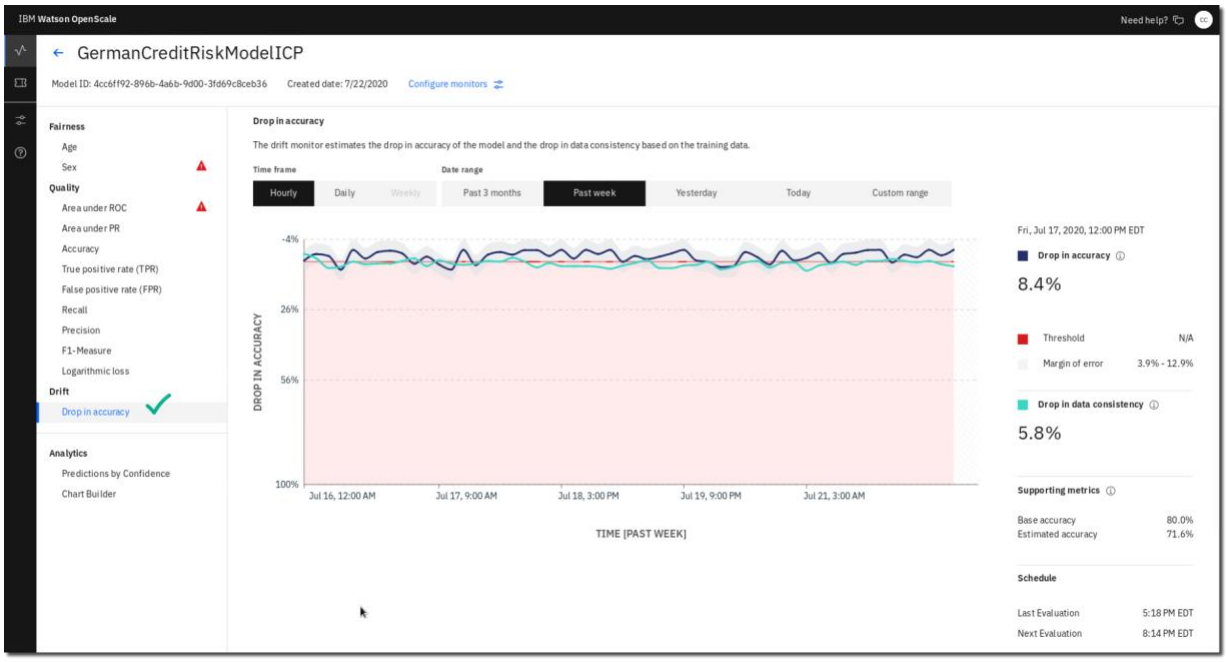


__33. The OpenScale drift monitor is a separate linear regression drift model, trained to determine which types of data the production model struggles to correctly predict. This drift model allows OpenScale to forecast potentially costly drops in model accuracy without requiring additional feedback data.

Additionally, the drift monitor compares incoming prediction requests with the training data to identify changes in data consistency that also may affect model output.

These two measurements are shown on the drift monitor screen. Estimated drop in accuracy is represented by the black line, drop in data consistency by the green line, and alert threshold by the red line.

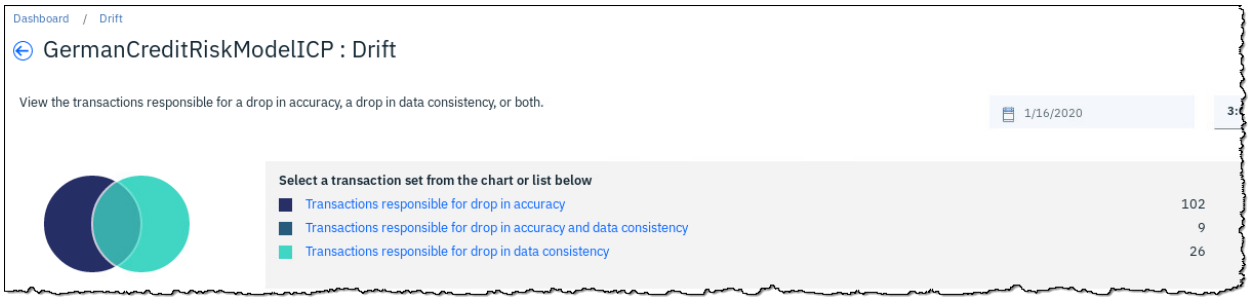
__34. Select the `Drop in accuracy` link.



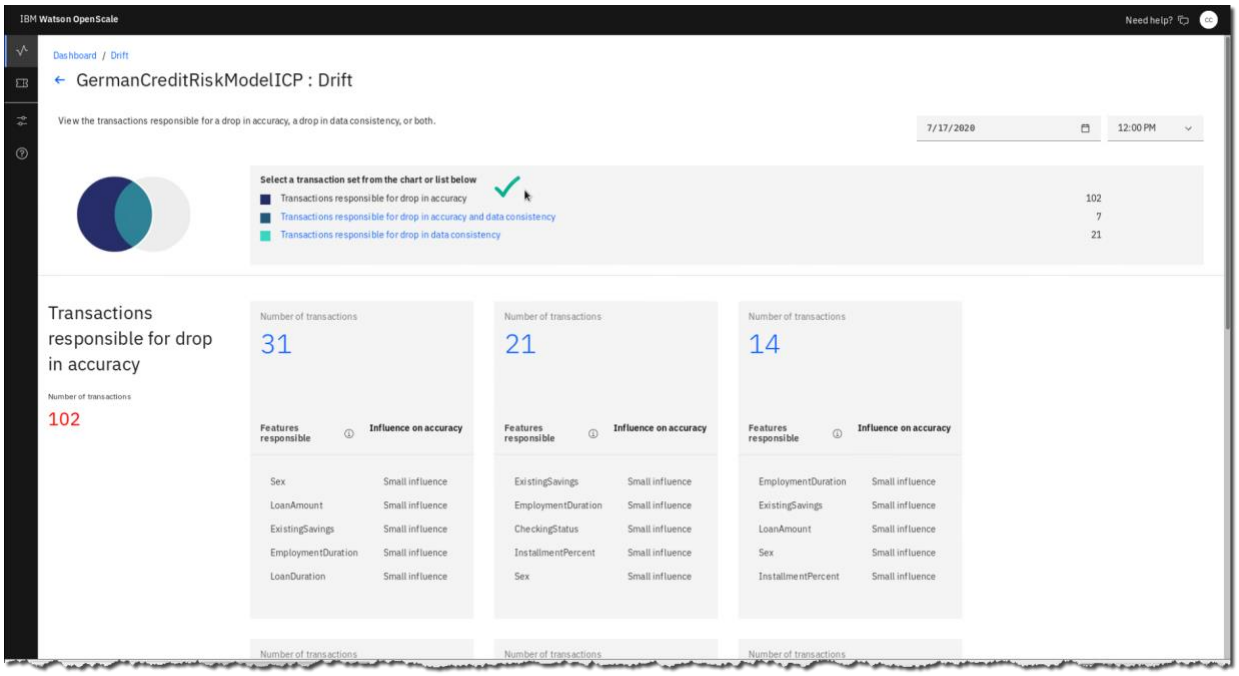
__35. Click the chart where the drop in accuracy is at its greatest `(8.4%)`.

NOTE: Your values may be slightly different than shown due to randomness in data used for training.

__36. Here, we can get a detailed view of the transactions responsible for estimated drops in accuracy, data consistency, or both.

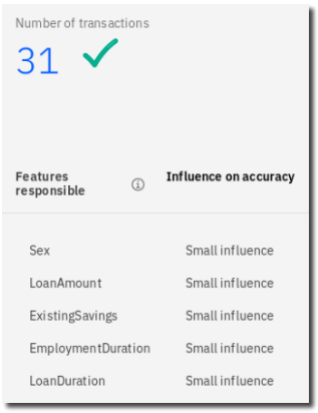


__37. Click `Transactions responsible for drop in accuracy.`



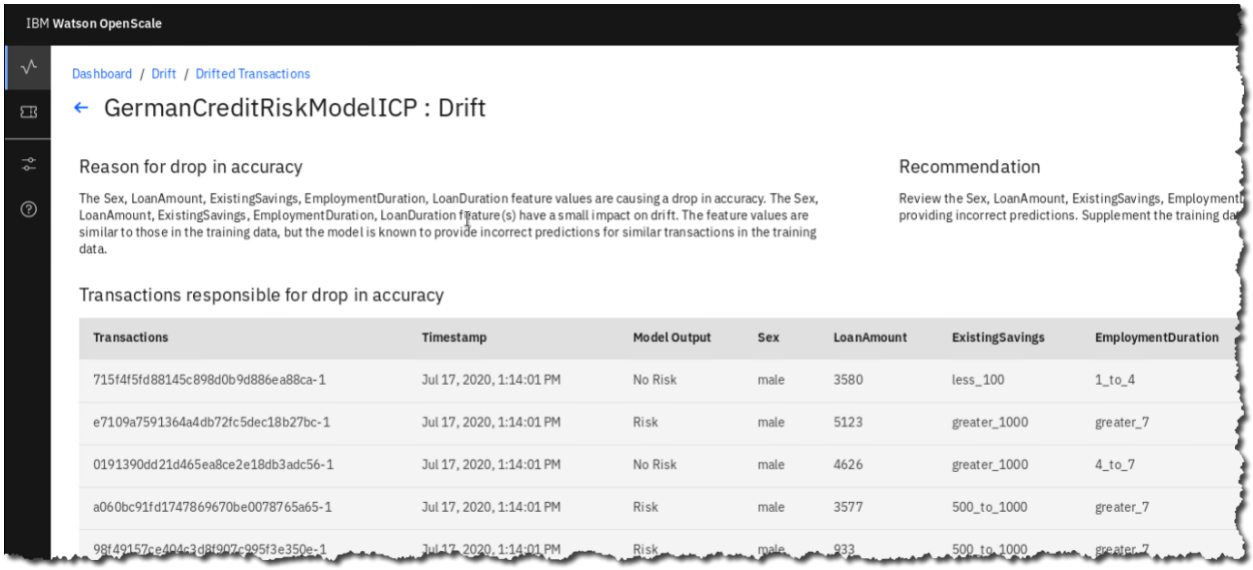
OpenScale divides transactions (predictions) that are affecting model accuracy into groups that share feature characteristics, providing a view of which feature values are causing our drift and how much influence each is having.

__38. Click on one of the tiles for a `transaction grouping`.



__39. Here, OpenScale provides a detailed summary of how the values are affecting the model, as well as recommendations for how to address the issues with corrected training data.

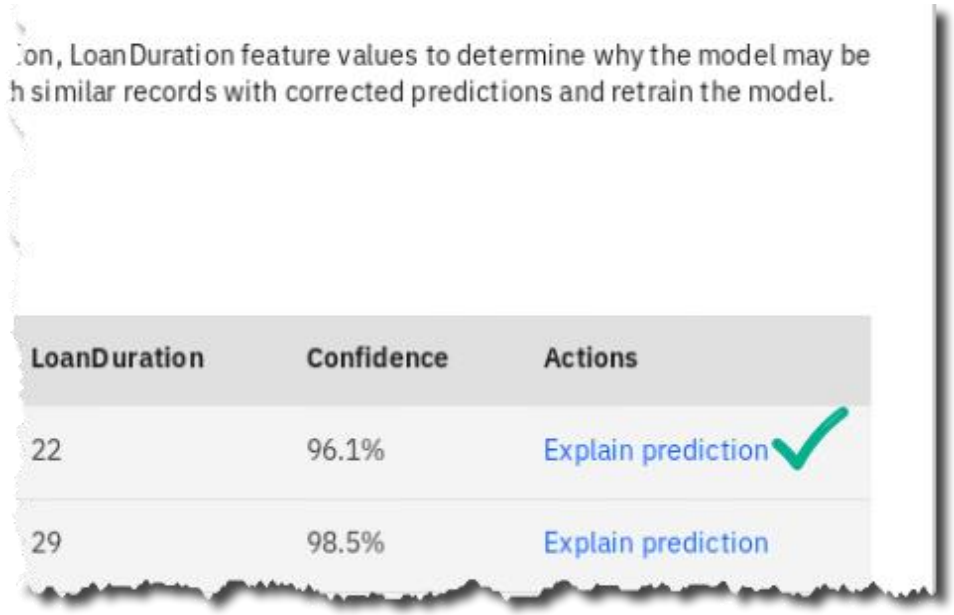
OpenScale lists the predictions and feature values for this grouping.



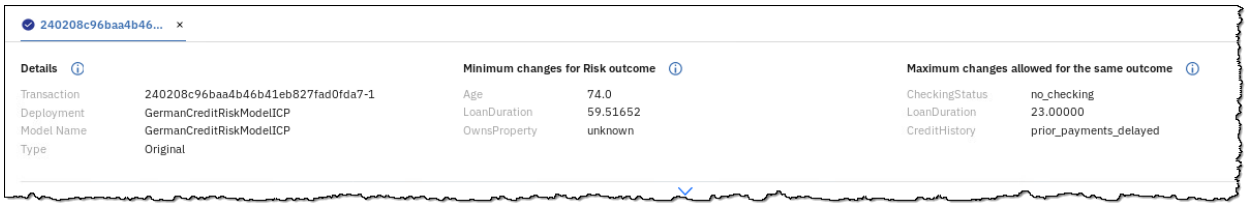
8.4.3 Explain an individual prediction

Using a variety of open source algorithms, OpenScale can provide highly detailed explanations of the predictions your model has made.

__40. Click the Explain prediction link in the Actions column of the transactions table.



__41. The OpenScale explanation feature works by slightly perturbing the feature values from the original prediction, sending these values to the production model, and measuring the impact the changes have on the outcome. By sending thousands of perturbed requests, OpenScale can gain a detailed picture of feature importance for not only relatively simple models like linear regression or decision tree classifiers, but also complex neural networks and image recognition models. (It can take a minute or two for the details to be calculated, so please be patient.)



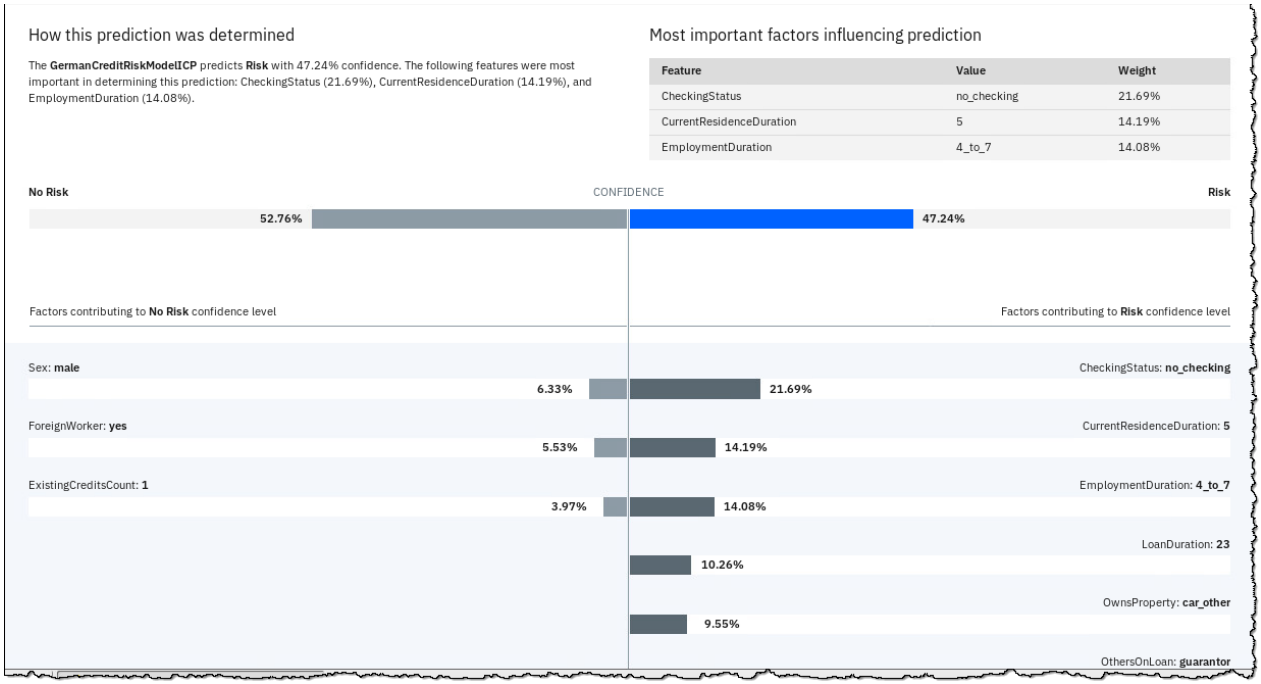
The upper portion of the screen shows information on the model and original prediction, as well as IBM's contrastive explanation technology. The Minimum changes for No Risk outcome show the Pertinent Negative values, or the least amount the feature values for the transaction can be changed to get a different outcome.

The Maximum changes allowed for the same outcome show the Pertinent Positive values, or how much the feature values can change and still have the model make the same prediction.

__42. Scroll down to see the factors that influenced the prediction.

Here, you can see the confidence the model has in its prediction as well as a quick summary of the most important factors that led the model to make a prediction that this particular loan application represents a risky one.

The chart at the bottom of the screen shows the feature values for this prediction, whether they contributed to a No Risk or Risk prediction, and how much they influenced the model.



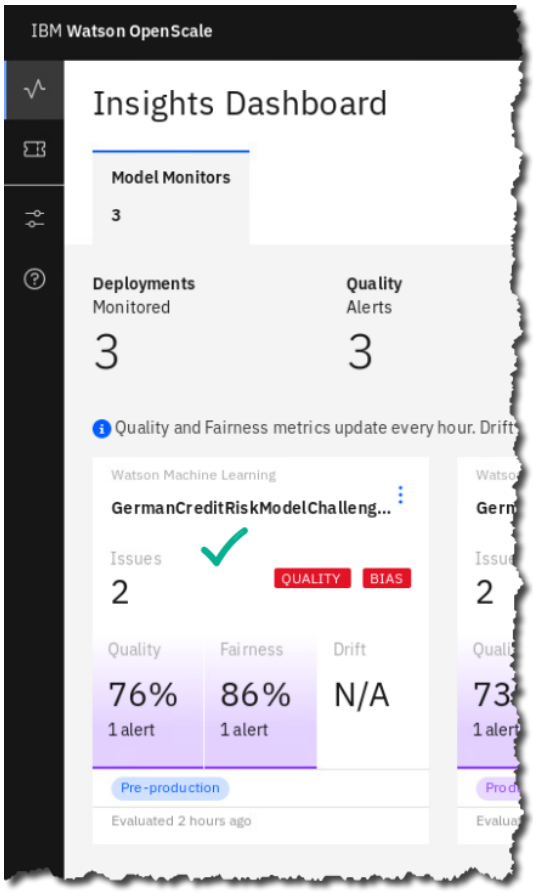
This detailed information allows you to ensure that your models are making predictions grounded in reality, as well as providing full explainability for predictions in case of an external audit or internal review of the model.

Finally, the data from this explanation is stored in the OpenScale data mart, where it, along with all other metrics, can be retrieved via API and surfaced to business users or even customers via dashboards or other applications.

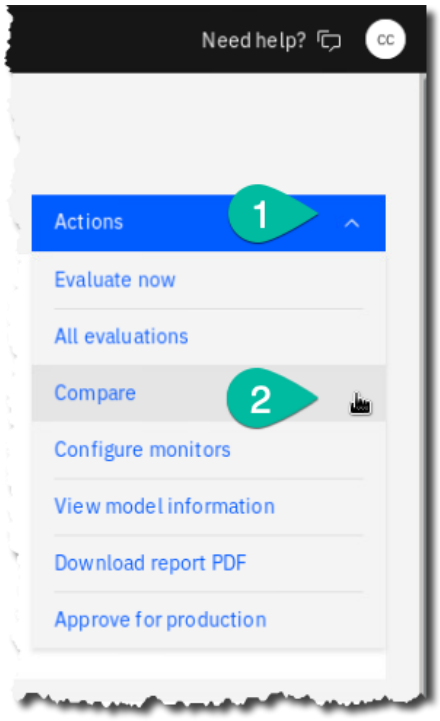
__43. Return to the Insights Dashboard ⇨ Model Monitors.

8.4.4 Comparing PreProduction Models

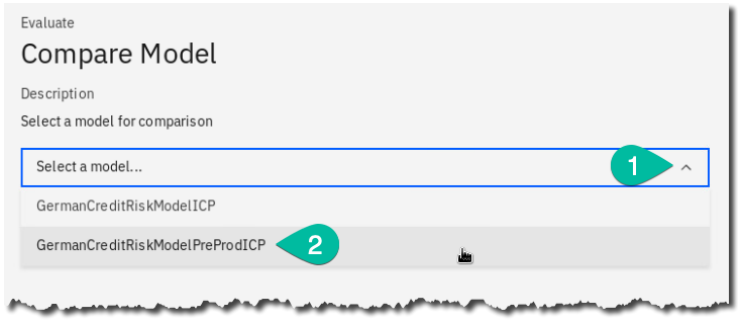
__44. You can also compare how pre-production models are performing. To do this, select the `GermanCreditRiskModelChallengerICP` box.



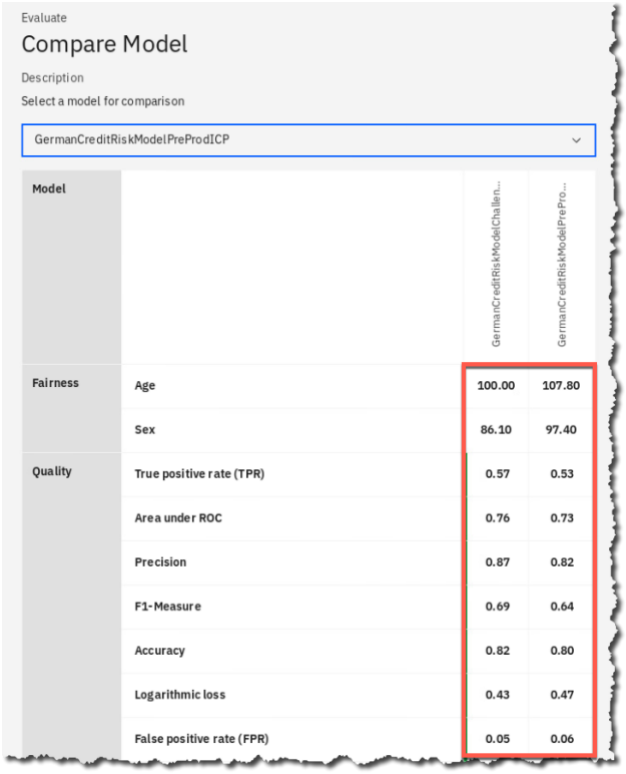
__45. Once opened, select `Actions - Compare` options from the list.



__46. Select the other pre-production model `GermanCreditRiskModelPreProdICP` from the list.



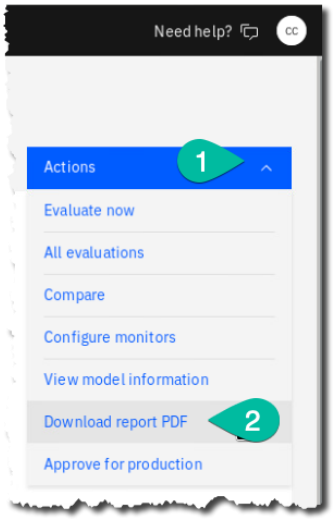
__47. Examine the comparative results of both models.



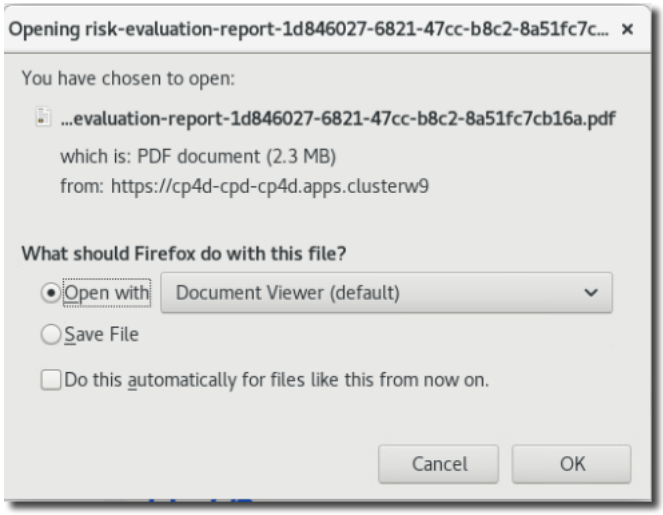
Model		GermanCreditRiskModelChallen...	GermanCreditRiskModelPrePro...
Fairness	Age	100.00	107.80
	Sex	86.10	97.40
Quality	True positive rate (TPR)	0.57	0.53
	Area under ROC	0.76	0.73
	Precision	0.87	0.82
	F1-Measure	0.69	0.64
	Accuracy	0.82	0.80
	Logarithmic loss	0.43	0.47
	False positive rate (FPR)	0.05	0.06

__48. Select the `X` to close the comparison.

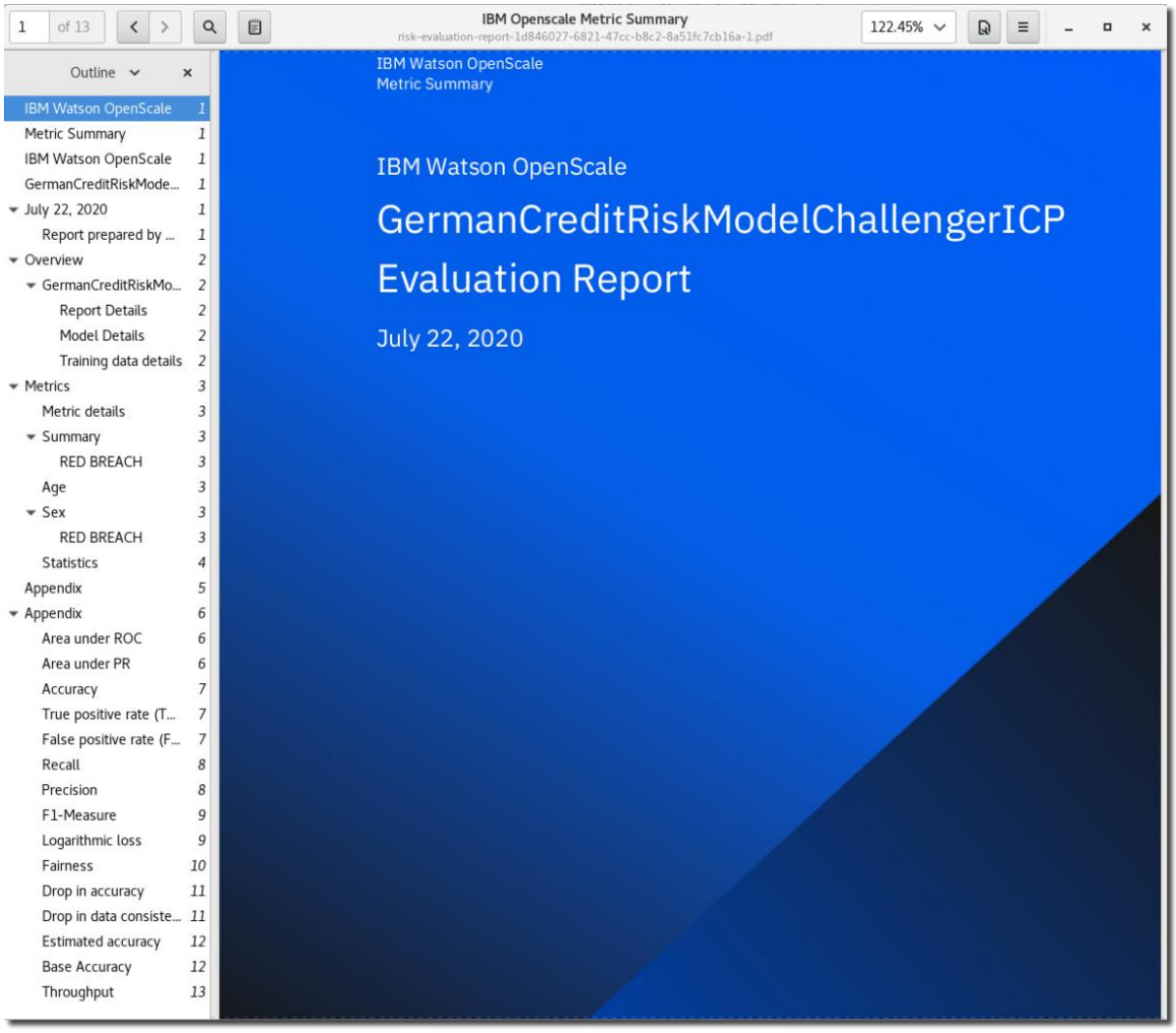
__49. One final step you can do is to deliver everything you have generated in a single PDF report. From the `Actions` menu, select `Download report PDF`. It may take a minute to prepare.



__50. If you are using the remote desktop experience, Select to `Open the document` with the default Document Viewer. If you are using a browser, simply download and open on your own machine.



__51. Browse through the report you have created!



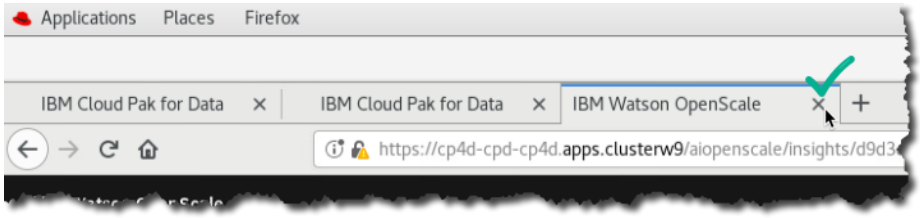
8.5 Monitoring the AutoAI model in OpenScale

Note: this section can only be completed as-is if you have previously completed the Analyze: AutoAI and Deploy labs.

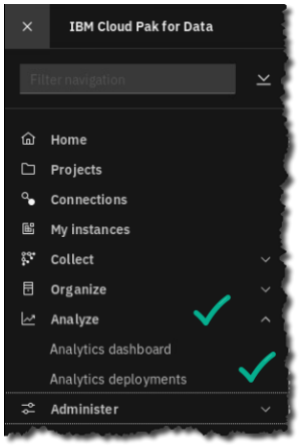
8.5.1 Associate the model with the new deployment space

Before you can monitor the AutoAI model in OpenScale that you created in a previous lab, you need to do a bit of housekeeping first to associate that model to the OpenScale deployment.

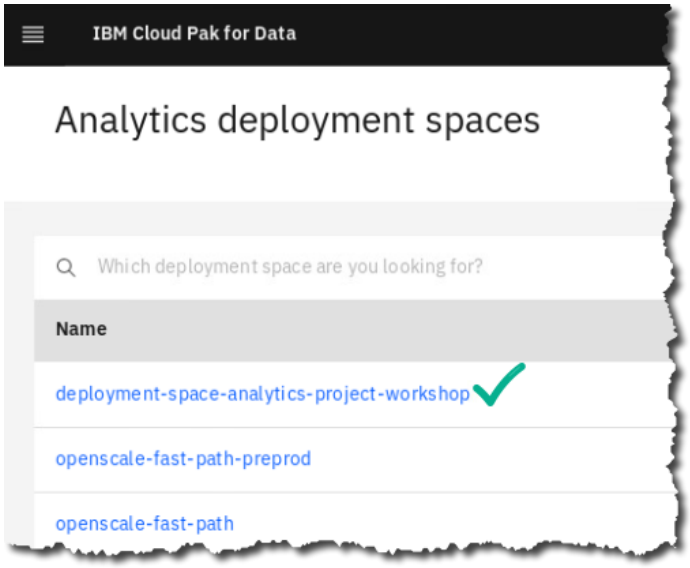
__52. `Close the current browser tab` and select your open `IBM Cloud Pak for Data` tab.



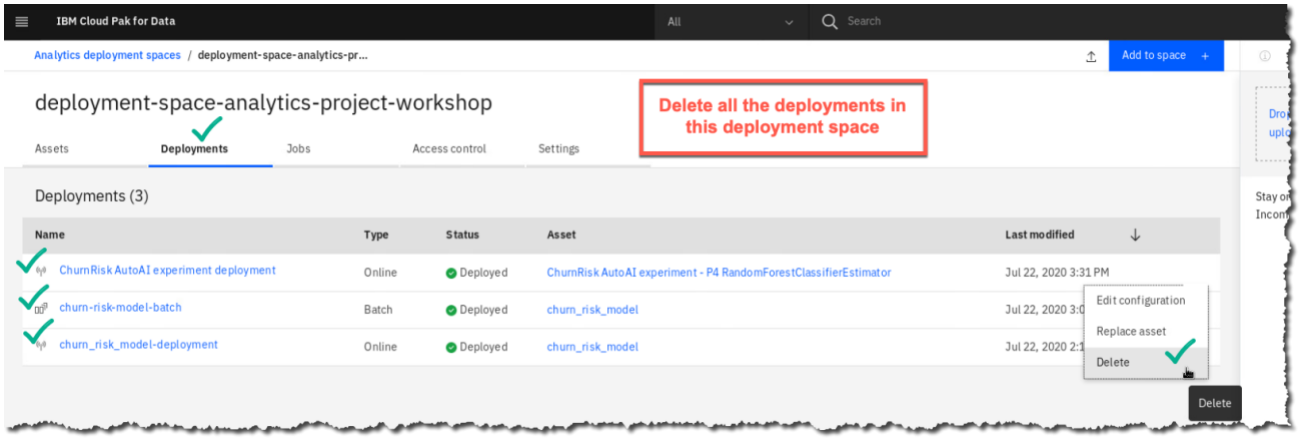
__53. Click `Navigation menu` ⇨ `Analyze` ⇨ `Analytics deployments,`



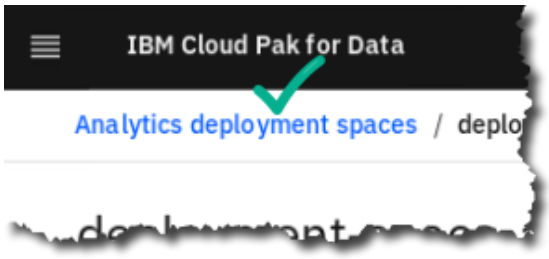
__54. Select deployment space `deployment-space-analytics-project-workshop.`



__55. Select `Deployments` tab, then select the `ellipses` next to every deployment and delete them all.

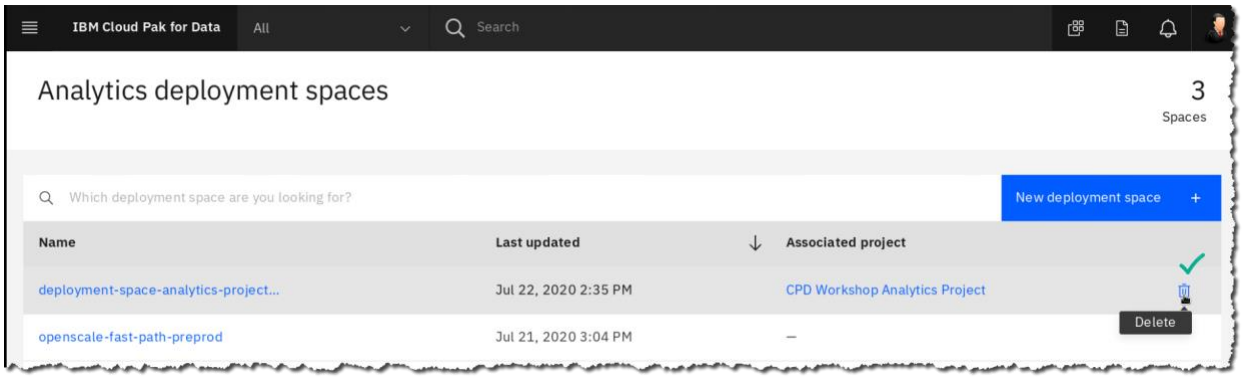


__56. Return to `Analytics deployment spaces` (using the breadcrumb on the page).

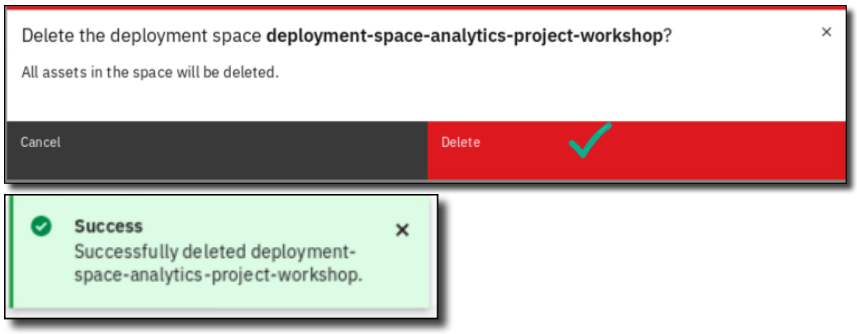


__57. With no deployment in it, you can now delete the deployment space `deployment-space-analytics-project-workshop.`

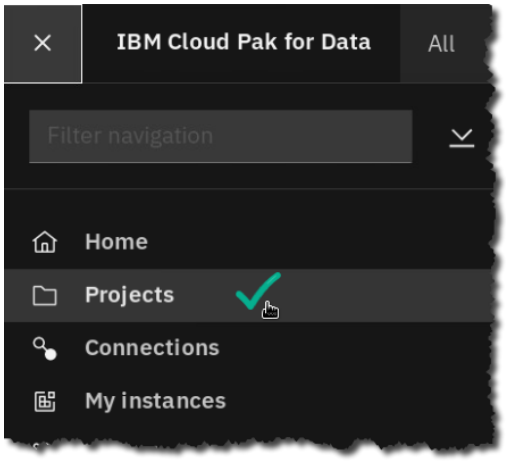
Click on the deployment, then `ellipses` ⇨ `Delete.`



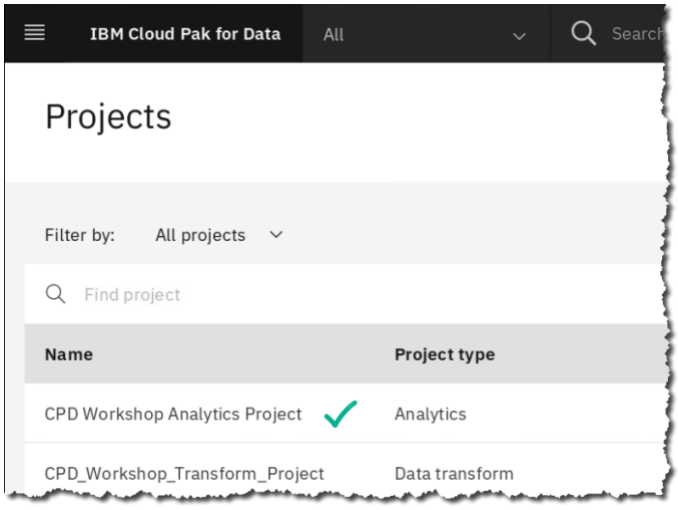
__58. Click `Delete` (to confirm). You should see the message as below.



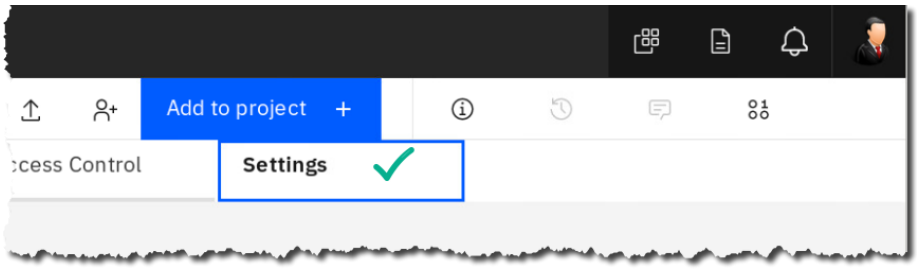
__59. Go to the `Navigation Menu` ⇨ `Projects.`



__60. Select the project: `CPD Workshop Analytics Project.`



__61. Click tab `Settings.`



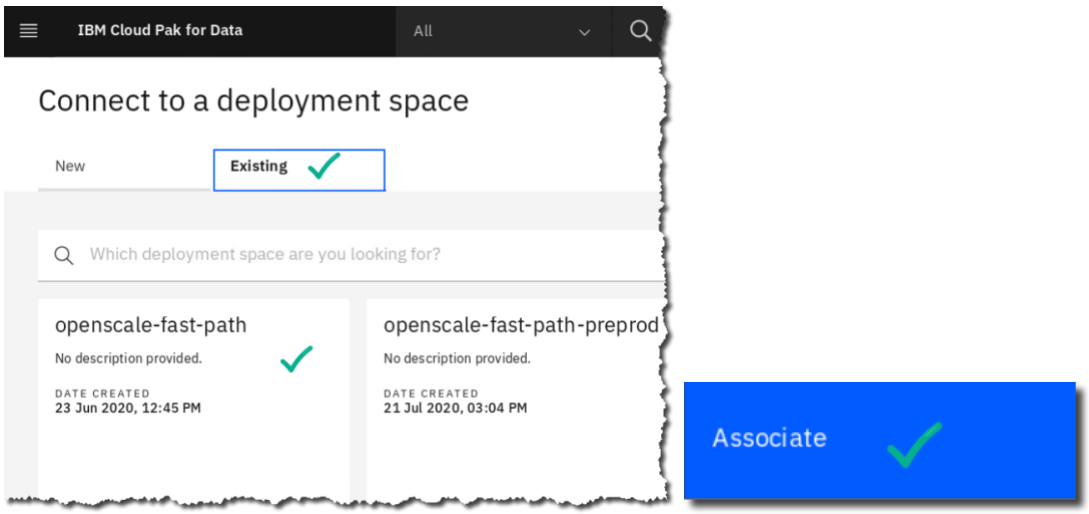
__62. Scroll down to find `Associated deployment space.`

Notice it says this project no longer is associated with a deployment space.

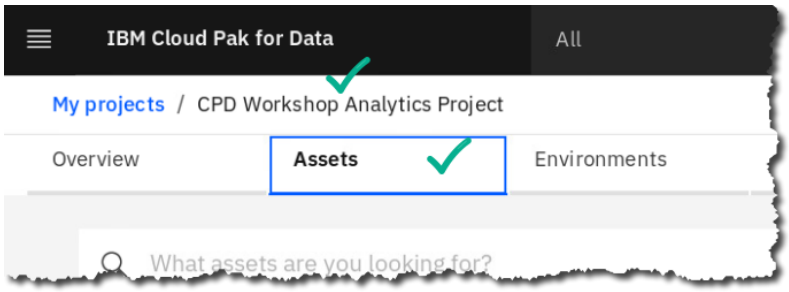
Click `Associate a deployment space.`

__63. Click section `Existing.`

Click on Deployment Space `openscale-fast-path` ⇨ `Associate.`



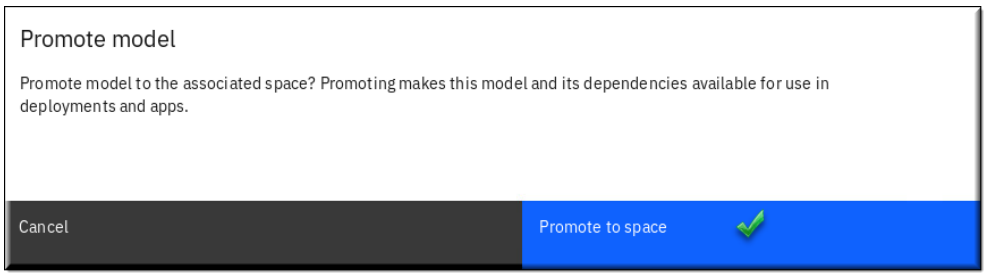
__64. While remaining in this project, go the tab `Assets,`



__65. Scroll to find section `Models.`

__66. Select the ellipses next to the model `ChurnRisk AutoAI Experiment – P4 RandomForestClassifierEstimator` – then click `Promote.`

__67. Choose `Promote to space.`

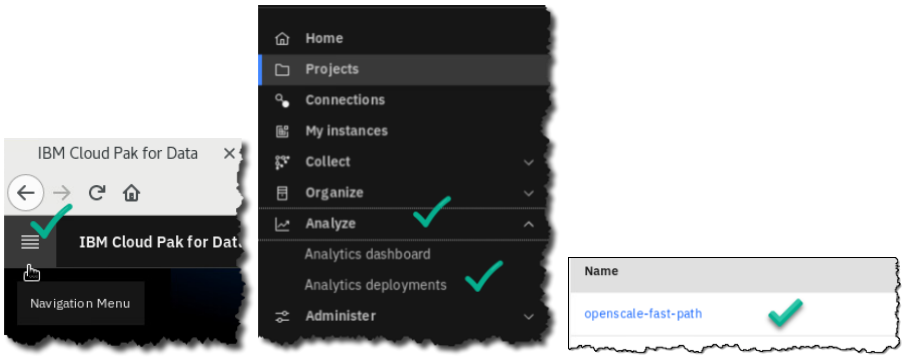


__68. The model is now promoted to the OpenScale deployment space the project is associated with.

8.5.2 Deploy the model in the new deployment space

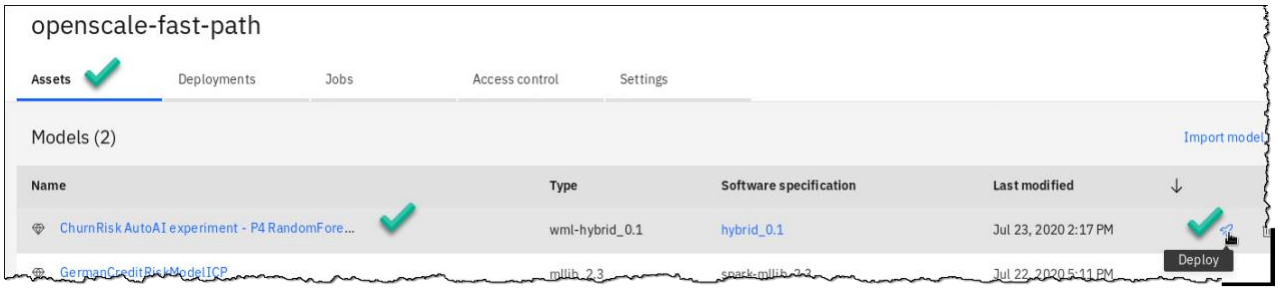
Now you can go to that deployment space to deploy the model.

__69. Click `Navigation menu` ⇨ `Analyze` ⇨ `Analytics deployments` ⇨ `openscale-fast-path.`



__70. Under the tab `Assets`, hover on the `launch` (rocket) icon for the model `ChurnRisk AutoAI experiment – P4 RandomForestClassifierEstimator.`

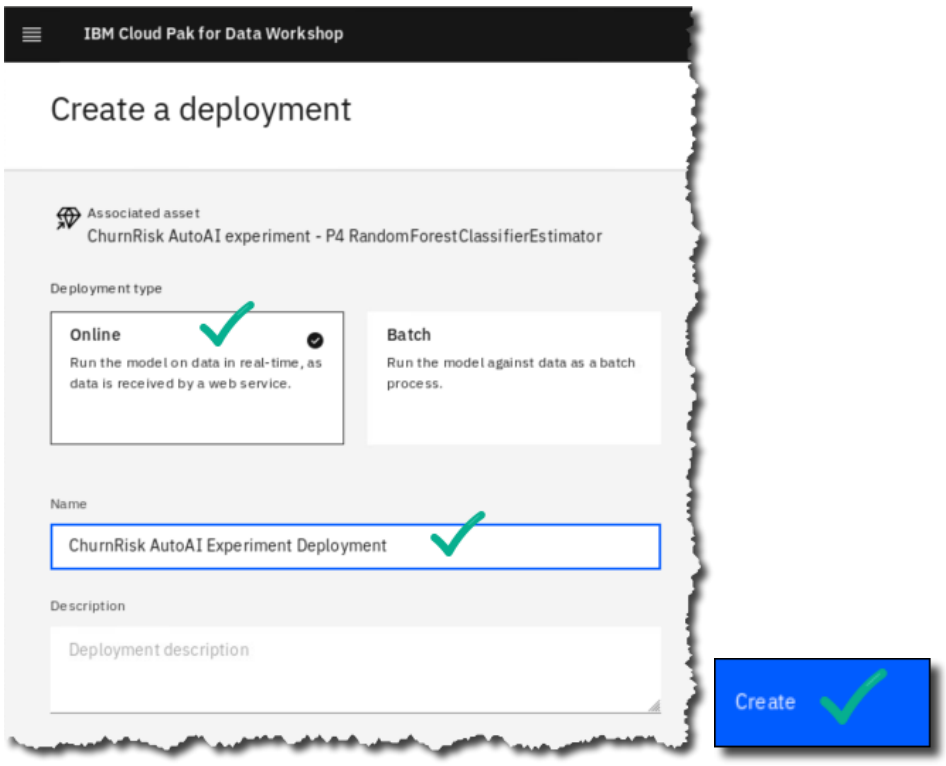
Click `Deploy.`



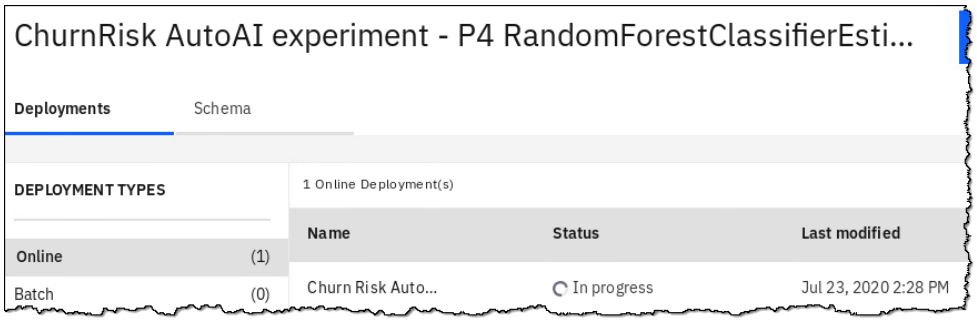
__71. Select `Online.`

Fill in Name: `ChurnRisk AutoAI Experiment Deployment.`

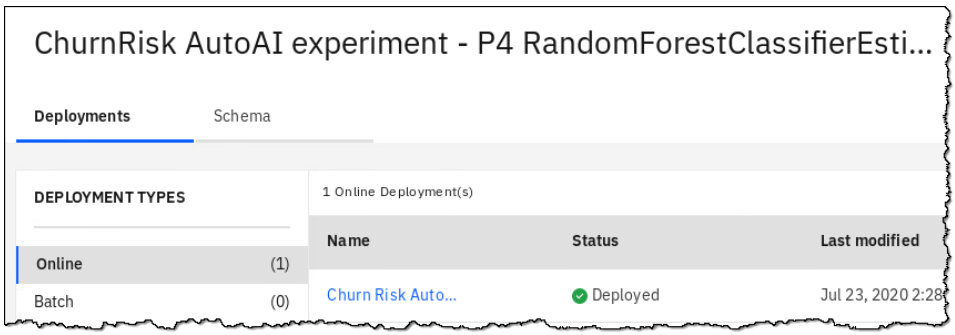
Click `Create.`



__72. You will see it as an `Online` deployment that is first: `In-progress.`



…in a minute or two, it will show as: `Deployed.`

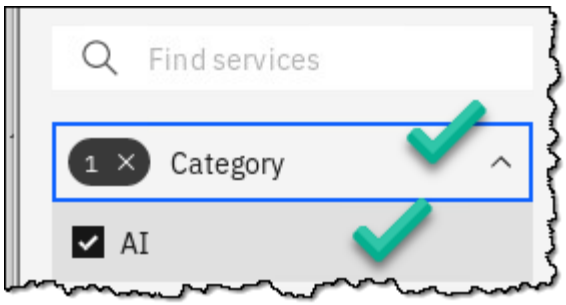


8.5.3 Monitor the model in OpenScale

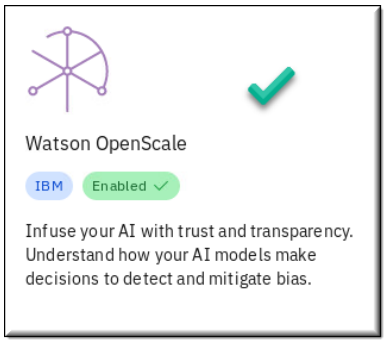
Now you will be able configure the monitoring of that model in Watson OpenScale.

__73. From the web client, click the icon for `Services` (at the top right corner of the screen).

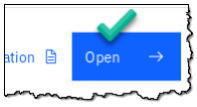
__74. Click `Services Category` ⇨ `AI.`



__75. Find and click on tile `Watson OpenScale.`

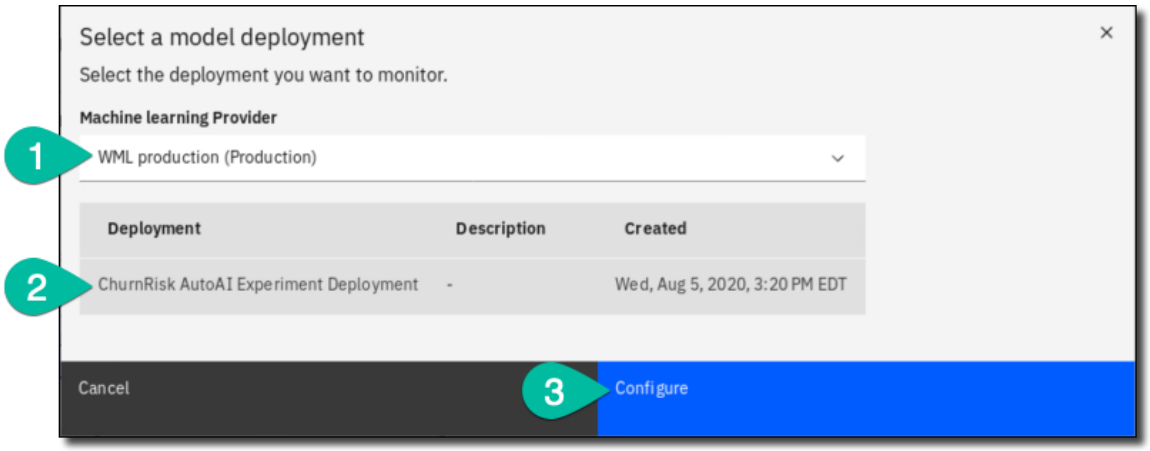


__76. Click (top right corner) `Open.`

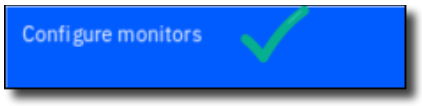


__77. Click (top right corner) `Add to dashboard.`

__78. Select the Deployment: `ChurnRisk AutoAI Experiment Deployment` ⇨ `Configure.`



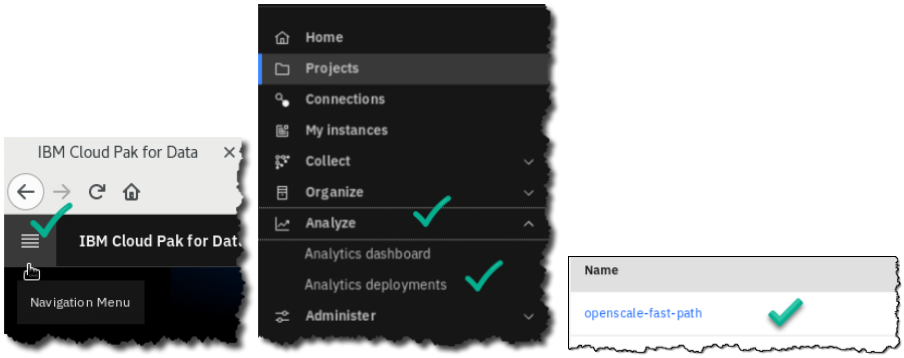
__79. Click `Configure monitors.`



__80. Next you will need to do at least one test (or payload log action) to activate the monitor.

If you are using the remote desktop, click the `Cloud Pak for Data web client` to open a new browser tab or duplicate your browser tab. If you are using your browser, duplicate the tab.

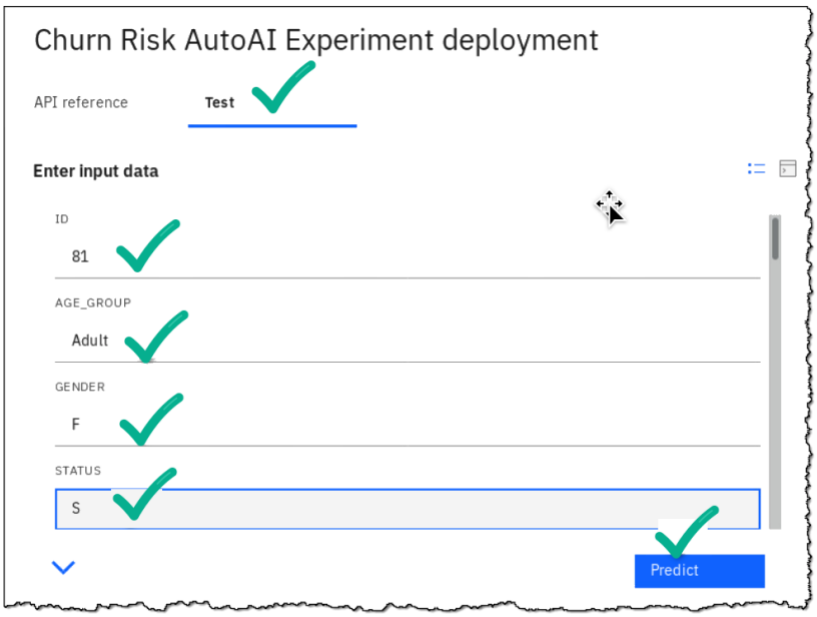
From that tab return to `deployment spaces` ⇨ `openscale-fast-path.`



__81. Choose deployment `ChurnRisk AutoAI Experiment Deployment.`

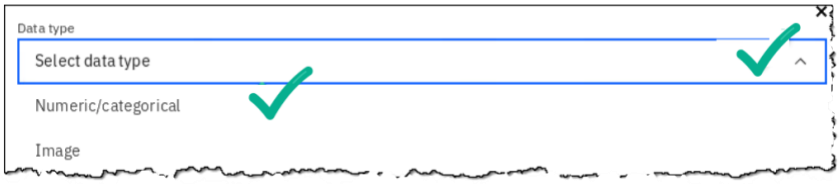
__82. Under tab `Test` enter the information shown below Click `Predict.`

(Note: you don't have to fill in all the features (input data) for this test, just the four values shown. This provides at least one payload log action… the prediction itself is not relevant here.)

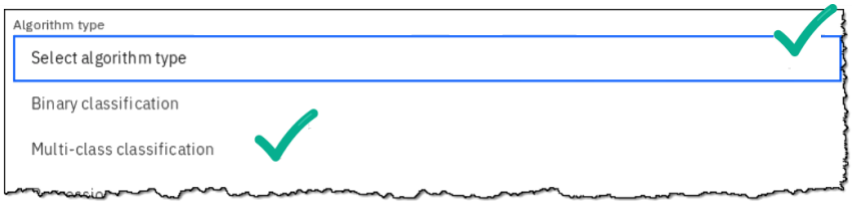


__83. Return to the OpenScale browser tab and click the `edit` (pencil) icon in the tile `Model Input.`

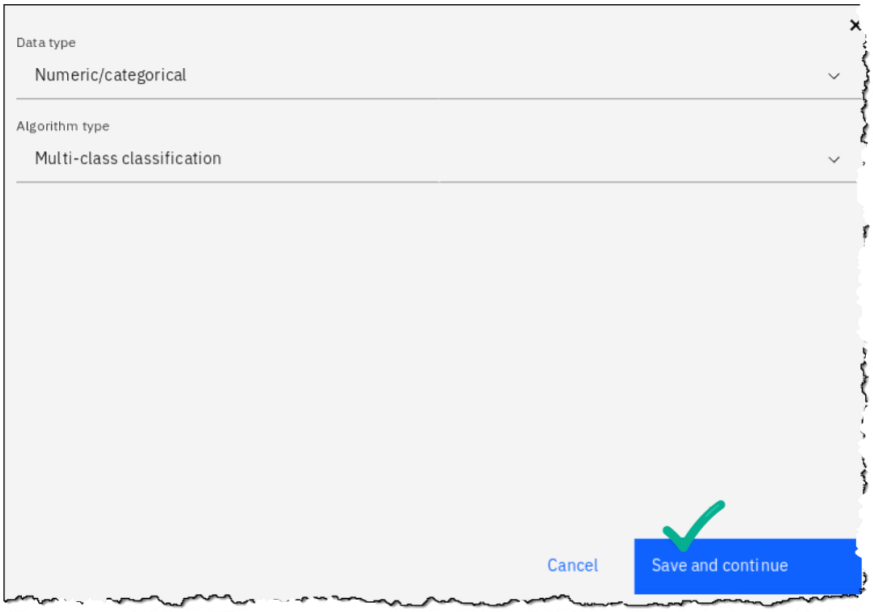
__84. Under `Select data type`, choose `Numerical/categorical.`



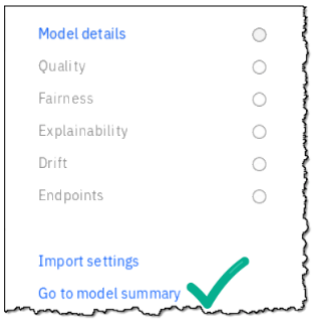
__85. Under `Select algorithm type`, choose `Multi-class classification.`



__86. Click `Save and continue.`



__87. Click `Go to model summary.`



For the remainder of the lab, try to complete the rest of the configuration tabs yourself.

The purpose of this exercise was to show you how to get started with model monitoring in Watson OpenScale with your own model. How to configure all of the options is probably best suited for a future Deep Dive lab.

8.6 Lab conclusion

Watson OpenScale is an enterprise-grade environment to help your organization infuse your AI with trust and transparency at scale, delivering a quick return on investment.

This Cloud Pak for Data service understands your models, detects and mitigates drift, and delivers explainable outcomes that are free from bias.

With this feature, Trade Co. is able to stay ahead of the curve.

**

Read here about how Watson OpenScale won a 451 Firestarter award in Q2 2019.

http://ibm.biz/openscale_award

**** End of Lab 08 – Infuse: Watson OpenScale**

Lab by Eric Martens, Burt Vialpando and Kent Rubin - IBM

Lab 09 INFUSE: COGNOS ANALYTICS - INTRODUCTION

9.1 Lab overview

In this lab, we will upload a CSV file given to us by our IT department to see if there is any interesting information and if we can use this information to better target market our potential high-risk churn customers.

For this lab, we will use the power of the IBM Cloud Pak for Data platform and the IBM Cognos Analytics cartridge. Installation of the IBM Cognos Analytics cartridge has already been done for you.

9.2 Persona represented in this lab

For this lab, we will assume the roles of both a Data Engineer and a Business Analyst.

Persona (Role)	Capabilities
 Business Analyst	Business Analysts deliver value by taking data, using it to answer questions, and communicating the results to help make better business decisions.

9.3 Logging into the CPD web client (if you have not already done so)

__1. If you are starting this lab stand-alone (without going through previous labs) do the following:

__2. Double-click the desktop icon: `Cloud Pak for Data Web Client.`

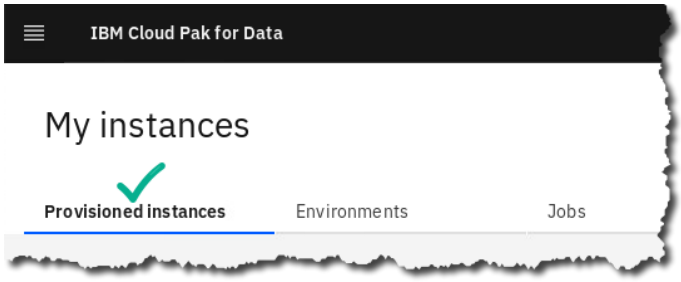
__3. The CPD web client GUI displays as shown. Use `cpduser` and `cpdaccess` for the *Username* and *Password* and click `Sign In`.

9.4 Logging into the IBM Cognos provisioned instance

__4. Select the `Navigation Menu` icon at the top of the screen.

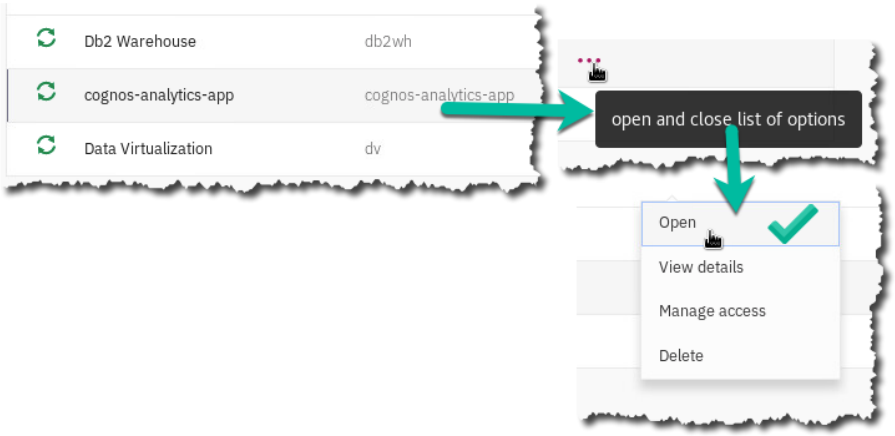
__5. Select `My Instances` from the menu.

__6. Select `Provisioned instances` from the tab at the top.



Here you will see a list of instances that have been provisioned for you already.

__7. Select the `ellipses` at the end of the `cognos-analytics-app` and select `Open`.



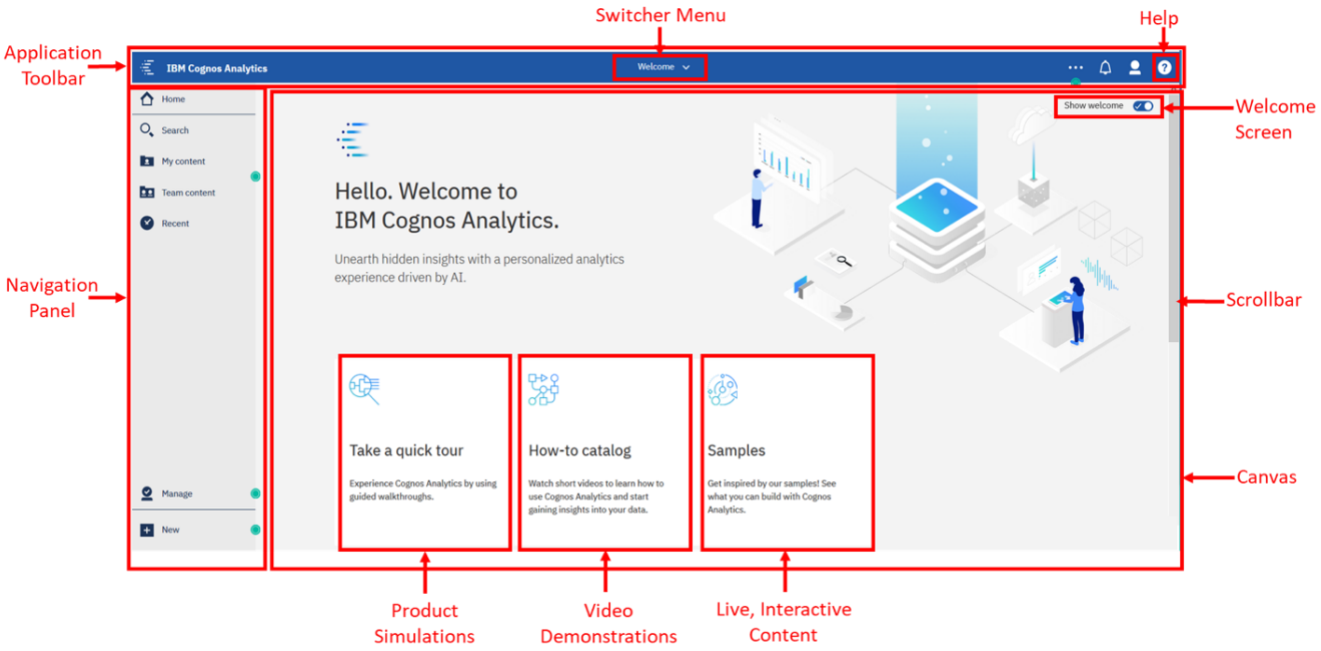
You will now be automatically directed and logged into the IBM Cognos Analytics instance on your IBM Cloud Pak for Data platform.

9.5 The IBM Cognos Analytics User Interface

The purpose of the User Interface (UI) is to provide Users with a streamlined way to get started using Cognos Analytics and view content and activities pertinent to them.

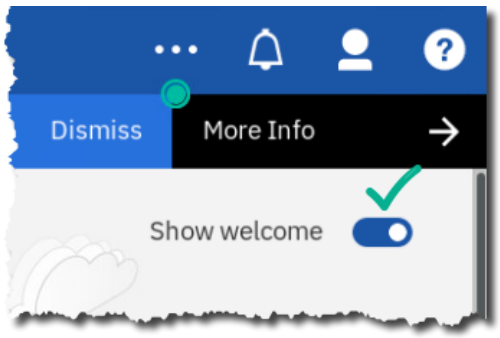
The User Experience brings you directly into the completely redesigned IBM Cognos Analytics User Interface (UI).

All IBM Cognos Analytics Users begin their navigation here at the `Welcome Screen.`

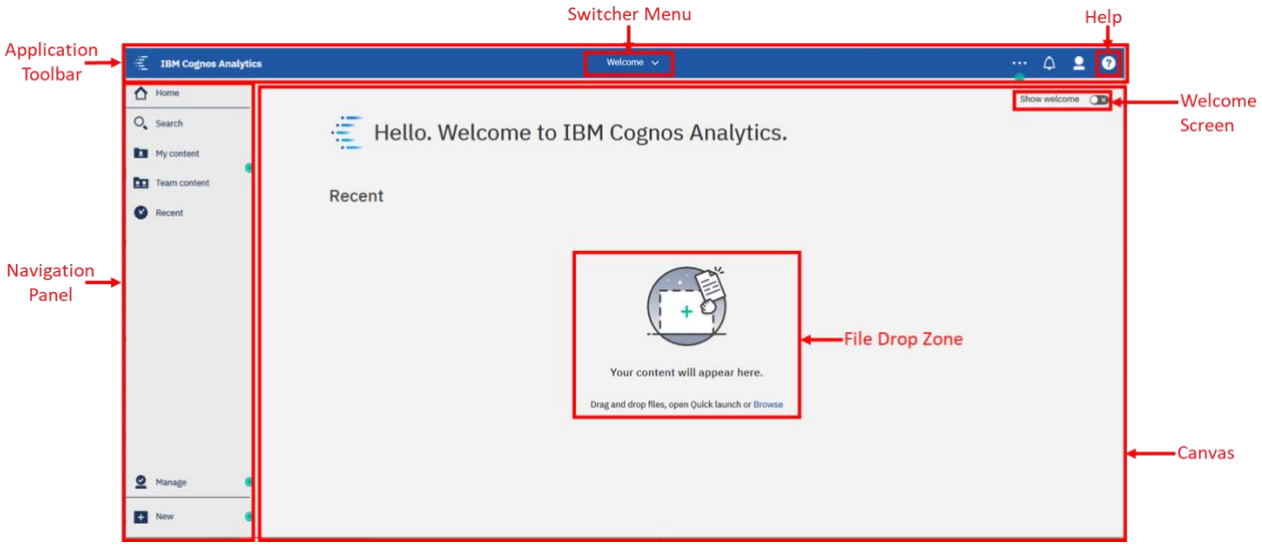


Business Analyst	TECH TIP: Not all `Welcome Screen Getting Started` tiles may be available in your workshop instance. The Welcome screen getting started tiles presented are based on which have been configured to run in the environment.

__8. Click on the `Welcome Screen toggle button` on the upper left of the canvas to collapse the Welcome Screen's Getting Started content.



__9. The canvas now shows the Recently used files, if any, in the **Recent** section, along with **the File drop zone** where users can easily upload their data files.

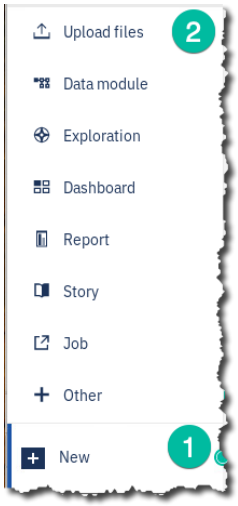


__10. Once you begin working with content, the canvas will update with your recently used items. In your Cognos Analytics instance, you may see recent content on the canvas.

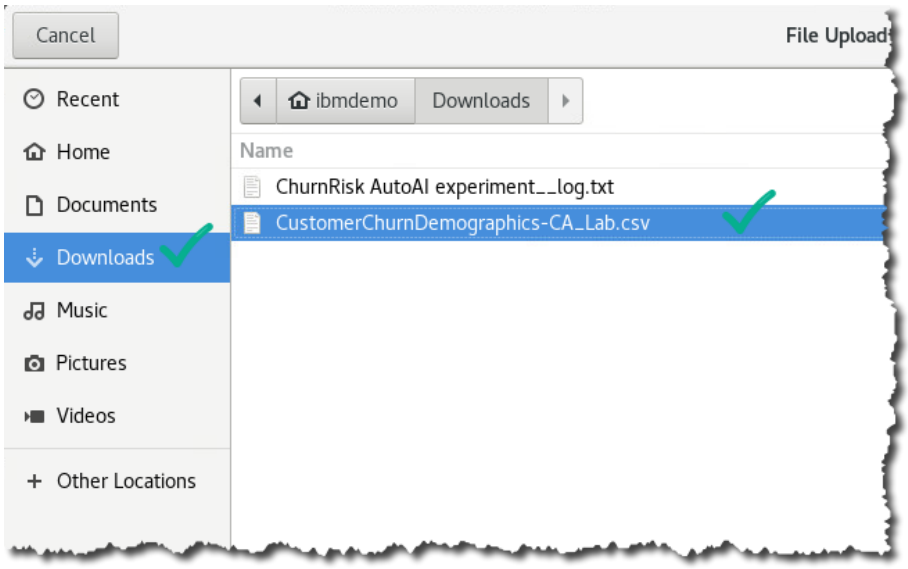
Now that we are in the IBM Cognos Analytics user interface, let's make a connection to the IBM Cloud Pak for Data – Virtualized Data. Remember, our business use case here is to analyze and create a dashboard quickly from this new set of data we virtualized. We did not have to wait for a Data Warehouse or Data Lake to be completed. Let's do some analytics at the speed of thought!

9.6 Importing the CSV file as an Exploration

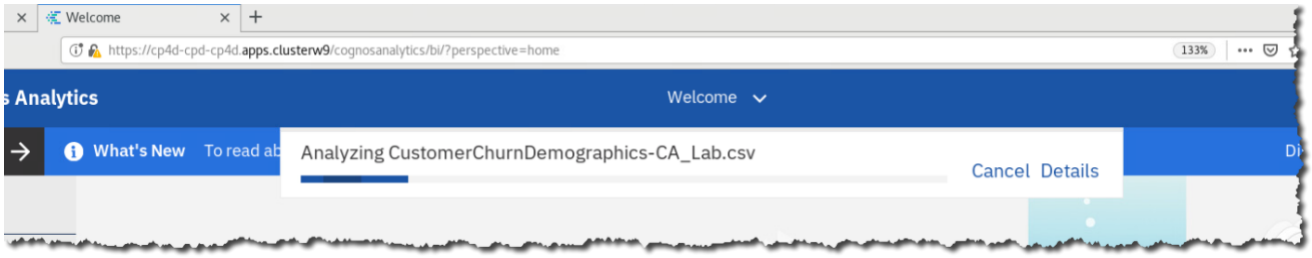
__11. From your Cognos Analytics Home screen, select `New – Upload.`



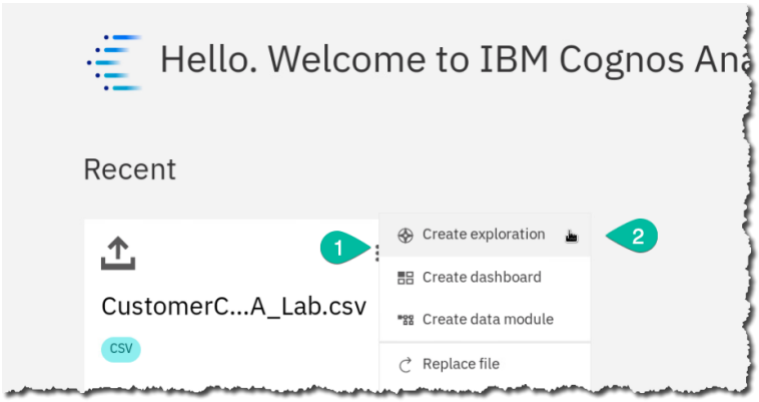
__12. Select `CustomerChurnDemographics-CA_Lab.csv` from `/home/ibmdemo/Downloads` if you are accessing via remote desktop. If you are accessing via browser, you may download the file from http://ibm.biz/Churn-Demographics.



__13. Cognos Analytics will then begin to import and analyze the data in the CSV file.



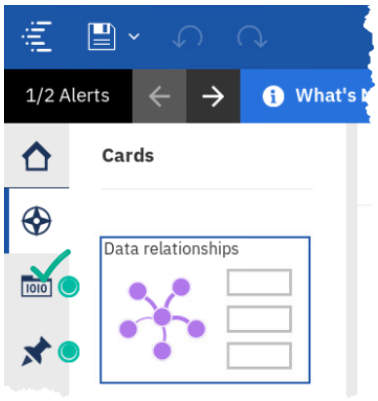
__14. In the Recent area, in the upper left corner of the CustomerChurn CSV card, select `ellipses` ⇨ `Create exploration.`



9.7 Cleaning up the data

When IBM Cognos Analytics imports CSV files, it does its best to understand the data type, usage and aggregation from reading your data. Depending upon the data you import, you may need to help IBM Cognos Analytics out by defining some data properties.

__15. From the left-hand menu, select the icon for the `Data tab`.



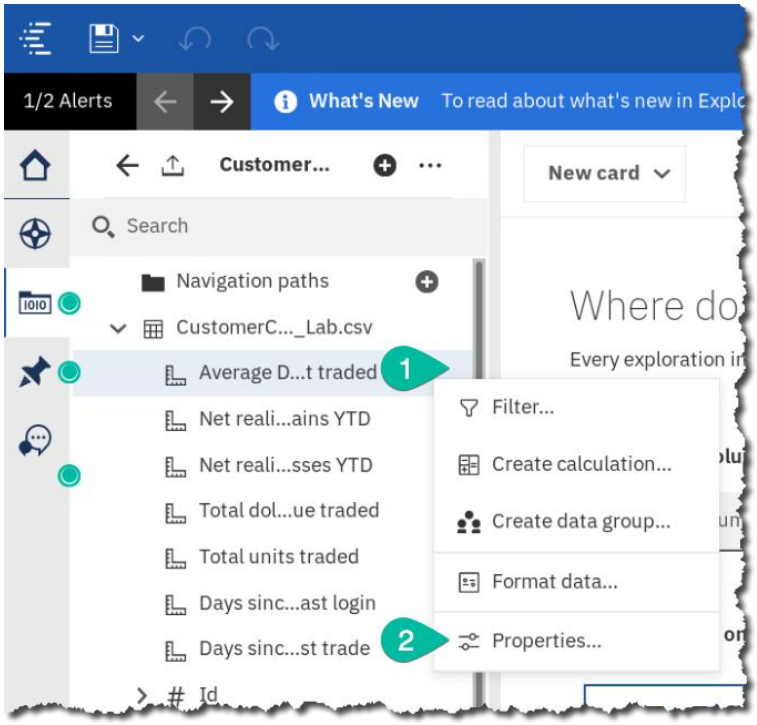
Note: You will notice icons next to each field showing a representation of the data items usage as recognized by IBM Cognos Analytics. Measures are represented by a ruler, numeric values as a hash, text values as ABC, time values as a clock and geographic points as a pin.

__16. Because this data is a snapshot total of values for a given date, you will need to change the aggregation properties of your measures, so they roll up correctly and do not give you misleading or false information.

 Business Analyst	TIP: If you cannot read the full name of the data item shown on your screen, simply hover over the name and a fly out will display the full name. Another option is to use the zoom out feature of your browser.

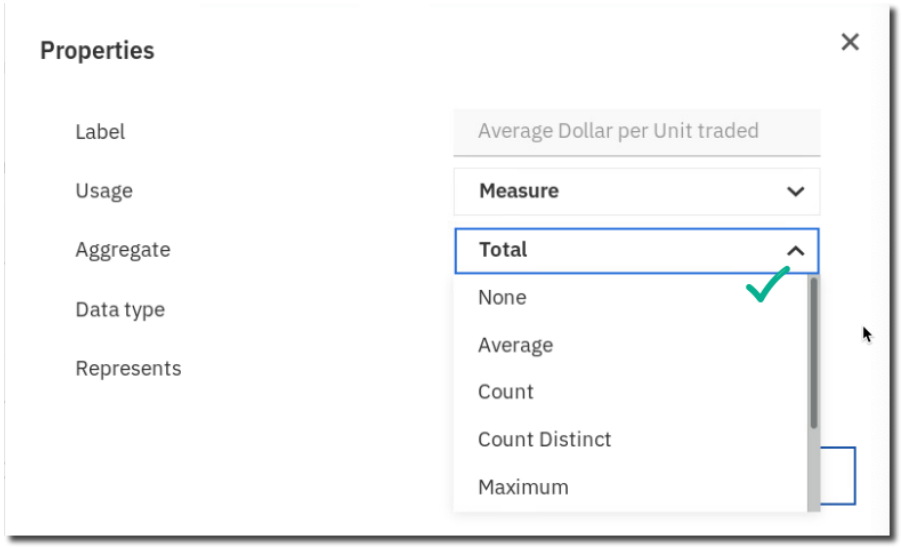
__17. For the following measures, change the Properties of the Data Item by selecting the `ellipses` at the end of each item and choosing `Properties` from the menu.

For example, the following displays the properties for `Average Dollar per Unit traded`:



__18. You will be presented with the `Properties` of that measure.

Change the `Aggregate` of `Average Dollar per Unit traded` measure to `None,` then `Close.`

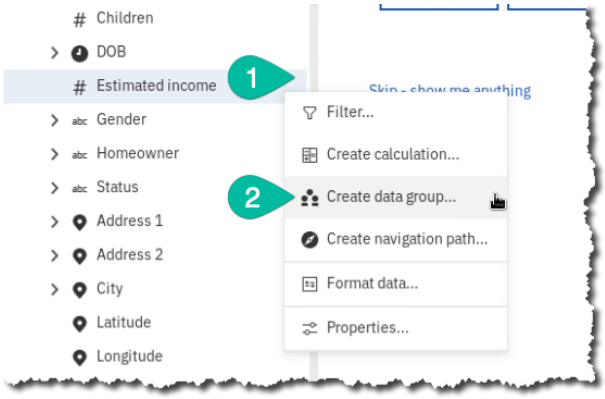


__19. Change the other `Data Items` with the following property values by following the previous steps.

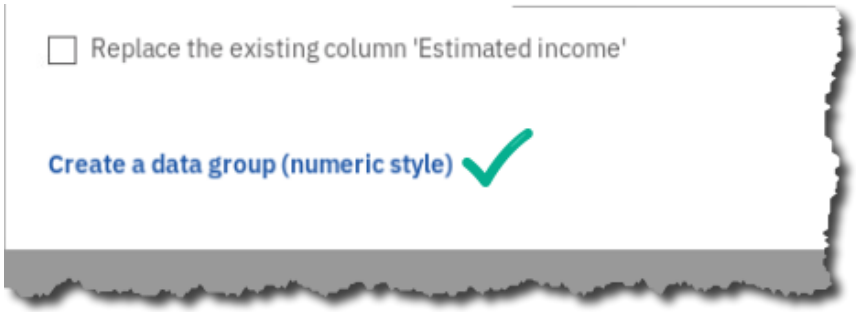
Data Item	Usage	Aggregate
Average Dollar per Unit traded	Measure	None
Net realized gains YTD	Measure	None
Net realized losses YTD	Measure	None
Days since last login	Attribute	Count Distinct
Days since last trade	Attribute	Count Distinct
Age	Attribute	Count Distinct
Children	Attribute	Count Distinct
Estimated Income	Attribute	None

__20. Because there are a lot of `Estimated Income` values, it would be better to group this data in buckets for easier understanding.

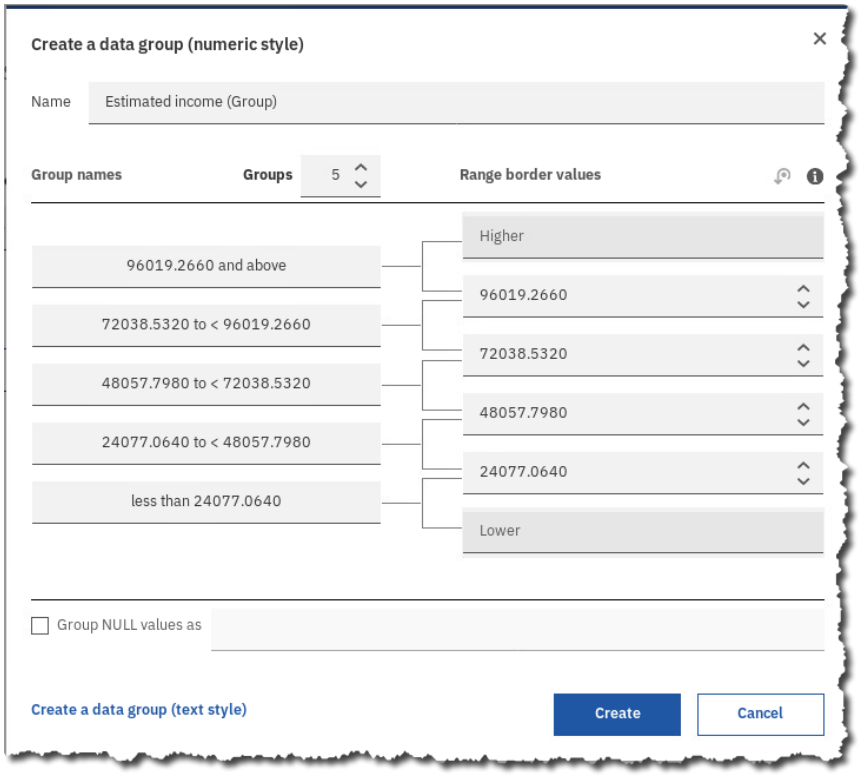
By `Estimated income`, select `ellipses` ⇨ `Create data group`.



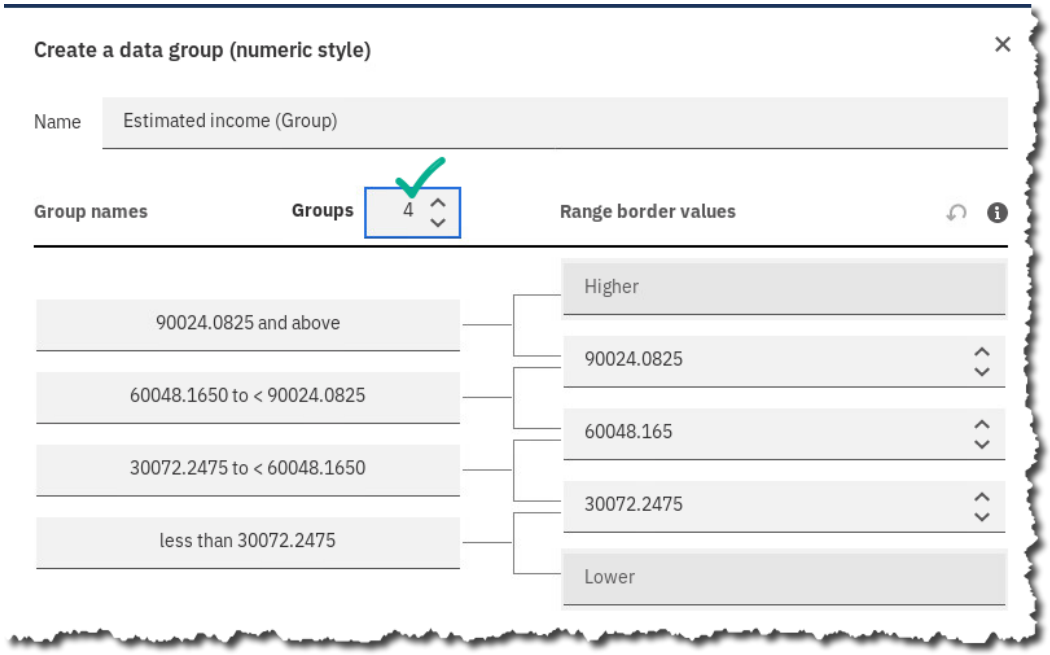
__21. Choose `Create a data group (numeric style).`



__22. You will be presented with the following auto grouping:



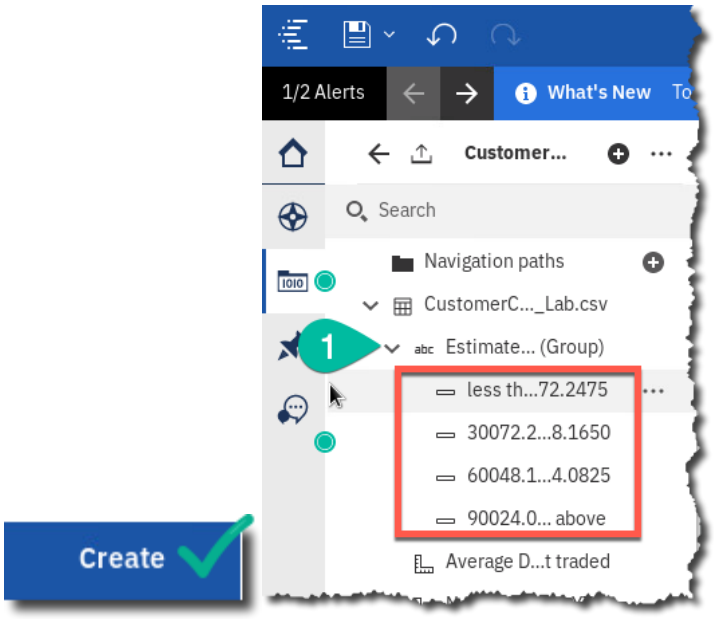
__23. Change the number of `Groups` to `4`.



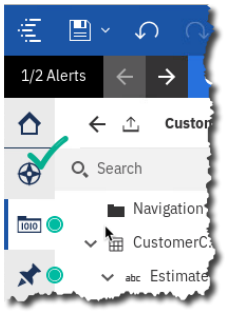
Note: You could manually enter group values or use Cognos to automatically do it for you!

__24. Select `Create.`

You will notice you have created a new grouping at the top of your Data Items list.



__25. When you are finished, return to the Exploration canvas by selecting the `Exploration` tab.



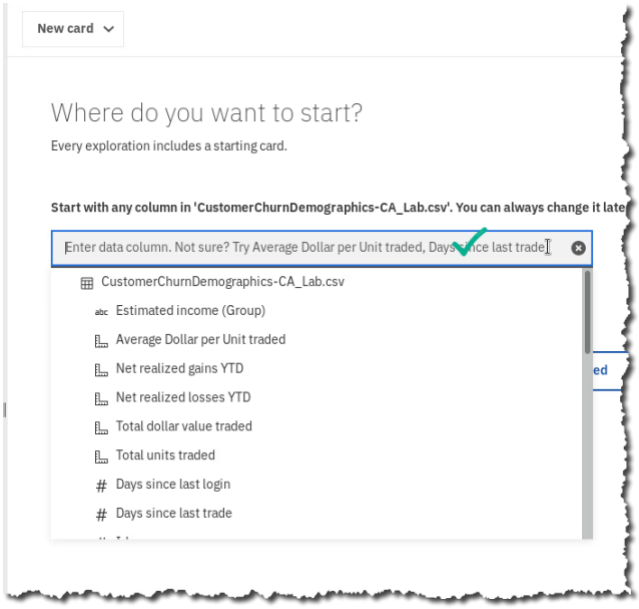
IBM Cognos Analytics Exploration asks you: `Where do you want to start?`

You can see that IBM Cognos Analytics suggests starting with some of the measures it found in your Data Module. This is helpful in discovering what are the possible drivers behind each of your metrics. IBM Cognos Analytics uses a powerful AI engine to assist you in discovering possible drivers between measures and dimensions, strength of relationships between dimensions, and also possible ways of examining this data for you by suggesting highly visual graphs of this data by dimension.

When IBM Cognos Analytics suggests: `Try starting with one of these…` it is providing you an easy way to begin discovering and exploring what it finds as potentially important (what is driving our metrics?) based on a measure or dimension.

We are not limited to these options, though.

__26. Select the area where it asks you to `Enter data column` and you will see that all of your Data Module Data Items are presented.



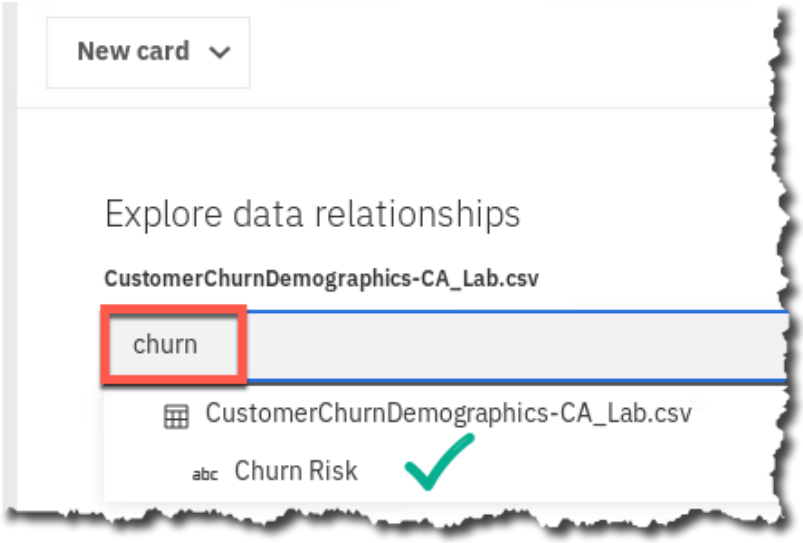
 Business Analyst	TIP: If you're really adventurous, or do not know the data you are exploring very well, you could select the `Skip – show me anything` link and IBM Cognos Analytics will select any measure as a starting point. Do not do this for this Workshop! It's good to know you can use the AI help from IBM Cognos Analytics when you need.

9.8 Exploring Data Relationships

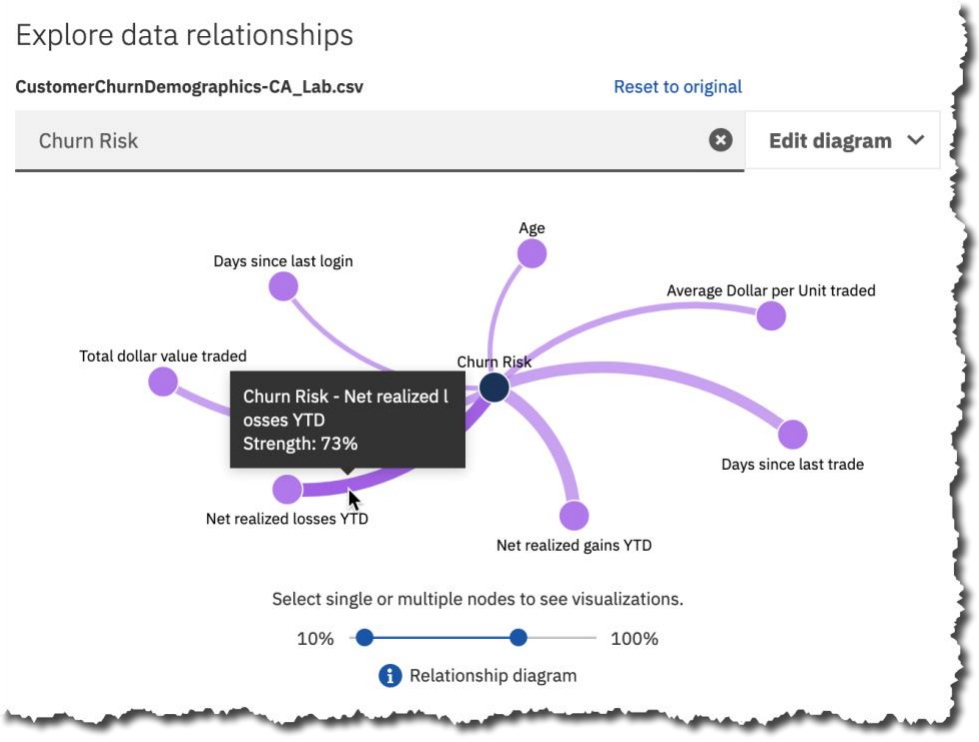
For this workshop, let's have IBM Cognos Analytics help us understand our customers better from this data set.

Let's first find out which customers are spending the most on trades with us.

__27. From the dropdown, type `Churn`, then select `Churn Risk.`



__28. By hovering over each of my measures associated with `Churn Risk`, I can get a quick understanding of relationship strength that is driving `Churn Risk`.

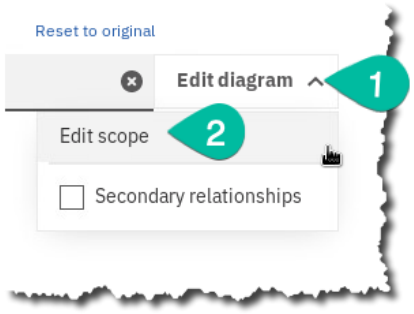


For example, there seems to be a 73% relationship strength between `Net realized losses YTD` and `Churn Risk`.

__29. Hover over other measures to see how they relate as well.

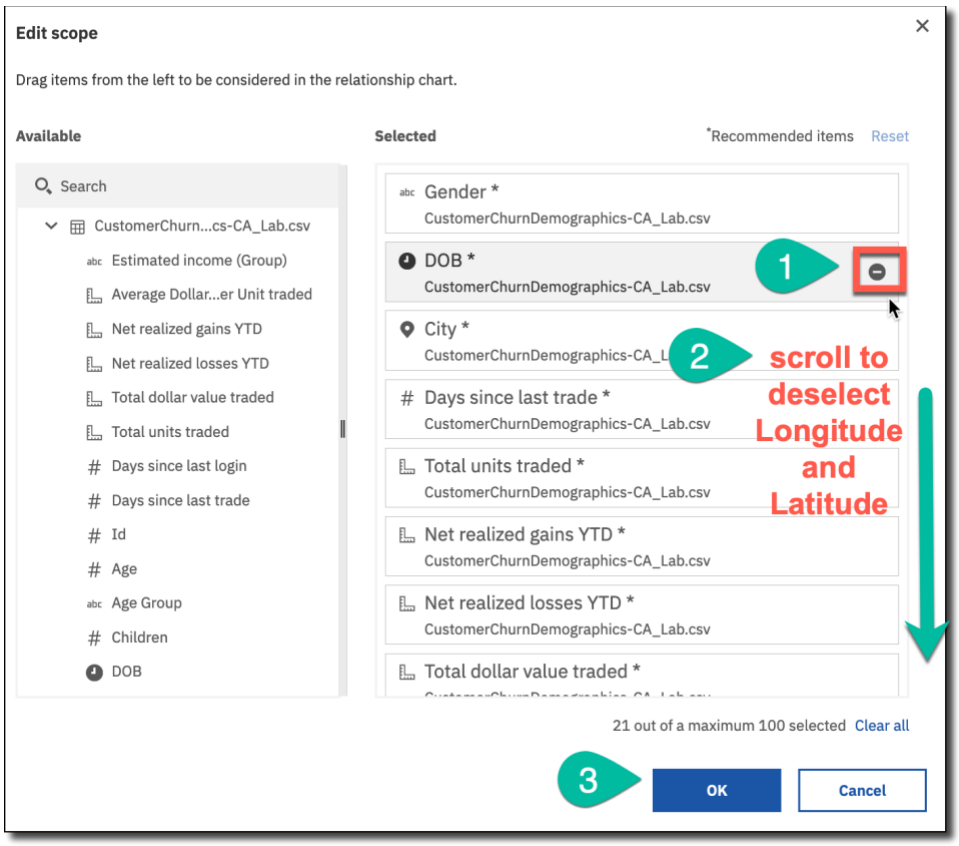
You can see that IBM Cognos Analytics has selected the scope of fields to show relationship strength between. You are not limited to this, however. You can change the scope of this relationship examination to remove fields that you know may interfere with getting a proper result.

__30. Select the drop-down menu to `Edit diagram` and select `Edit scope.`



Fields such as `DOB`, `Longitude` and `Latitude` are probably Data Items that do not bring any value when examining your data relationships and may cause static when exploring your data.

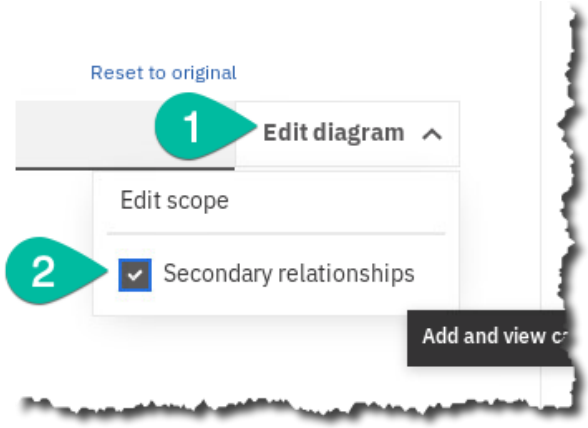
__31. You can remove those by hovering over, then selecting the ⊖ next to `DOB, Longitude` and `Latitude,` then `OK.`



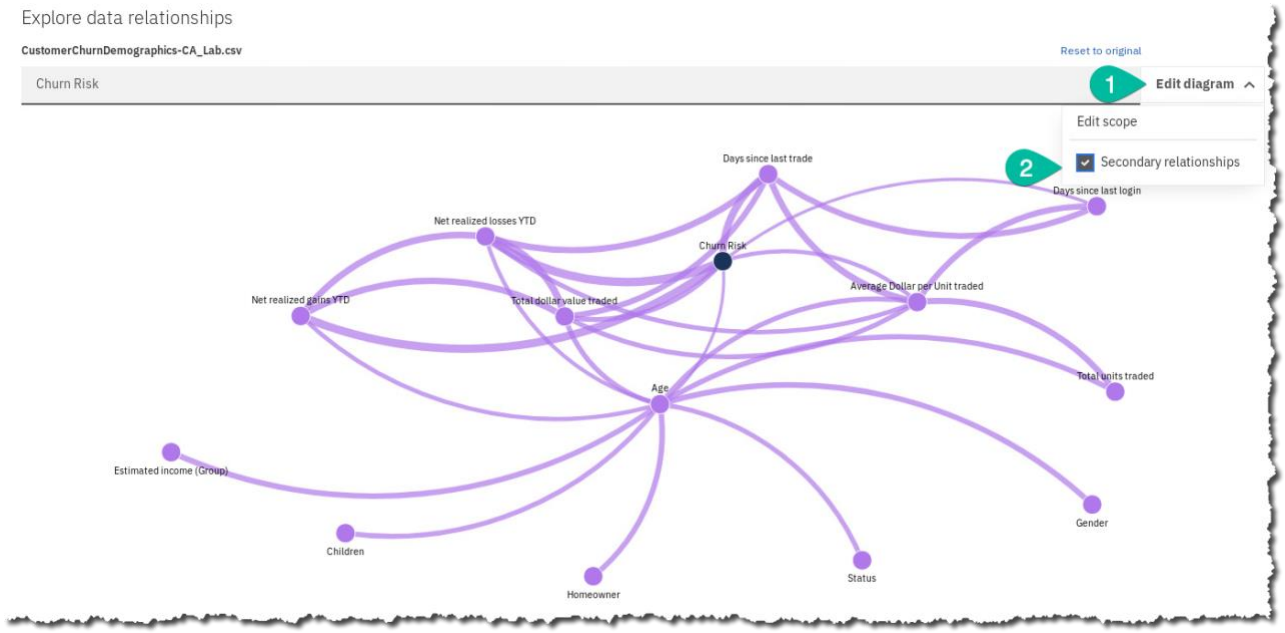
IBM Cognos Analytics gives you the flexibility to remove fields that you may not require when exploring your data and may interfere (static) with your analysis.

__32. You can also investigate secondary relationships to see the strength of the relationship between all fields.

Select `Edit diagram` ⇨ `Secondary relationships`.



__33. Hover around each of the fields to see view the relationship strength IBM Cognos Analytics discovered between them.

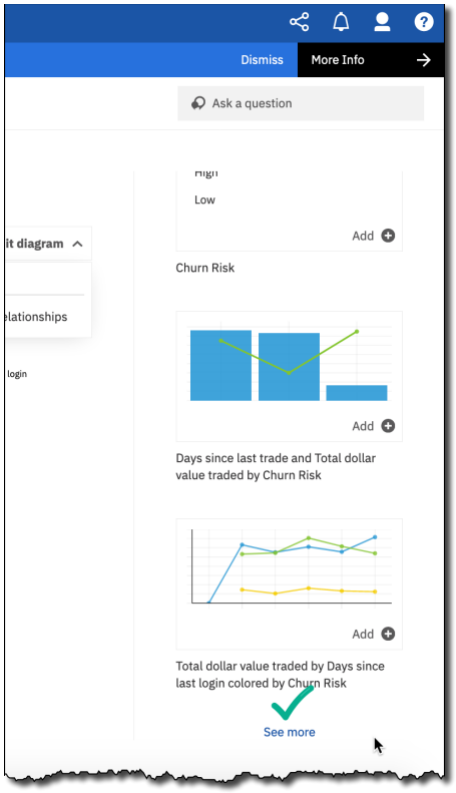


9.9 Creating Exploration Cards

When IBM Cognos Analytics examines relationships, the powerful AI engine will begin to surface charts and graphs that may be of interest to you. This is especially helpful if you are unfamiliar with the data you are working with.

`Scroll down on the far-right portion of your screen`, then notice that IBM Cognos Analytics has presented you with three visualizations. You can begin with one of these info graphs or 'Cards' or see more Cards.

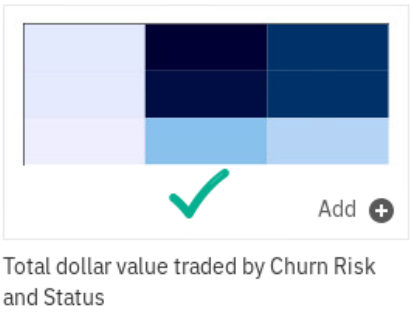
__34. Select link `See more`.



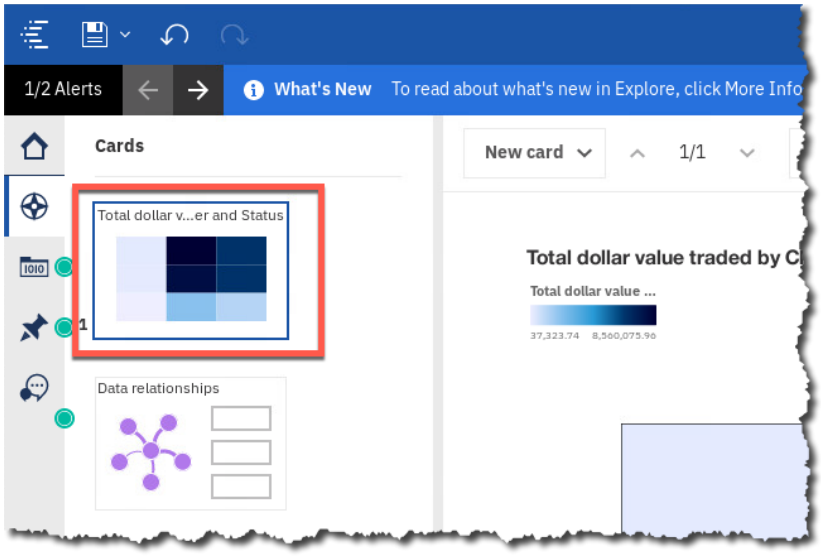
__35. You are now able to see more auto-generated cards via the power of IBM Cognos Analytics AI.

IBM Cognos Analytics is using its powerful AI engine to discover new possible drivers or interesting information from your data source.

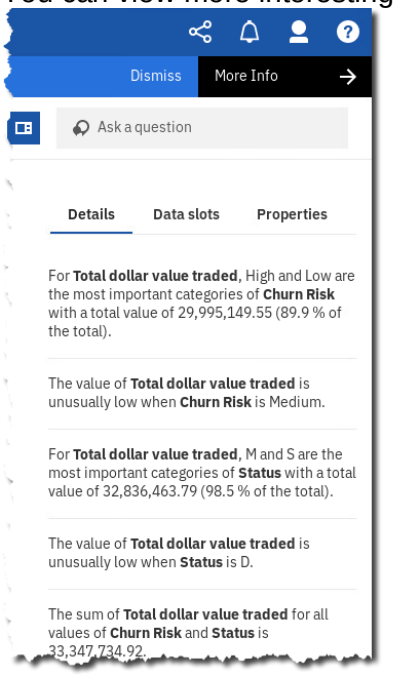
__36. Select inside the card: `Total dollar value traded by Churn Risk and Status.`



__37. Notice that it has added the Card to the left side in the Exploration Cards area.

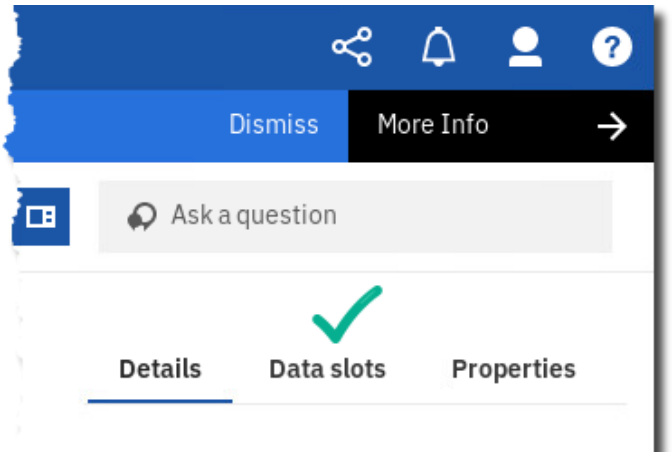


__38. We can see that our High and Medium Churn Risk customers are mostly married. You can view more interesting discoveries by viewing the `Details` tab.



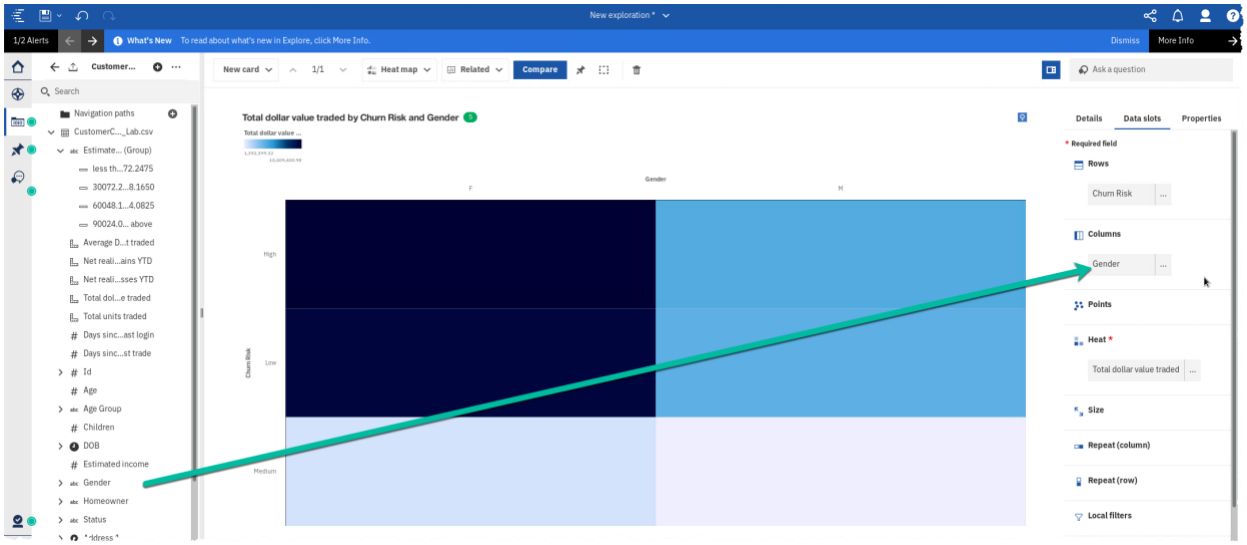
Let's now see if Gender plays a role in how much money they trade with us.

__39. Select the Data slots tab at the top next to Details.

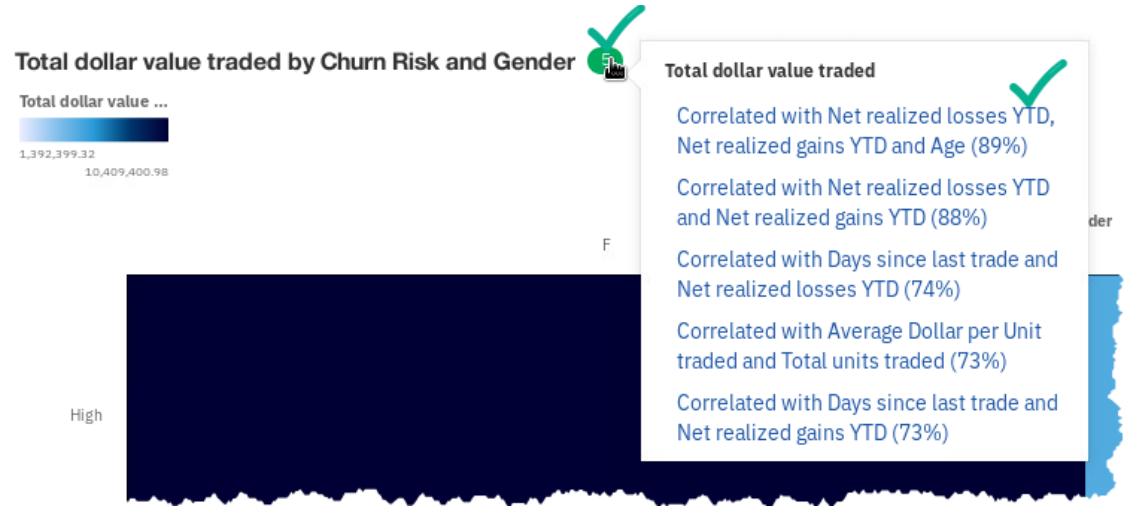


__40. Cognos Analytics provides you interesting insight to your data automatically, but you can take control, too. Select the data tab on the far-right side.

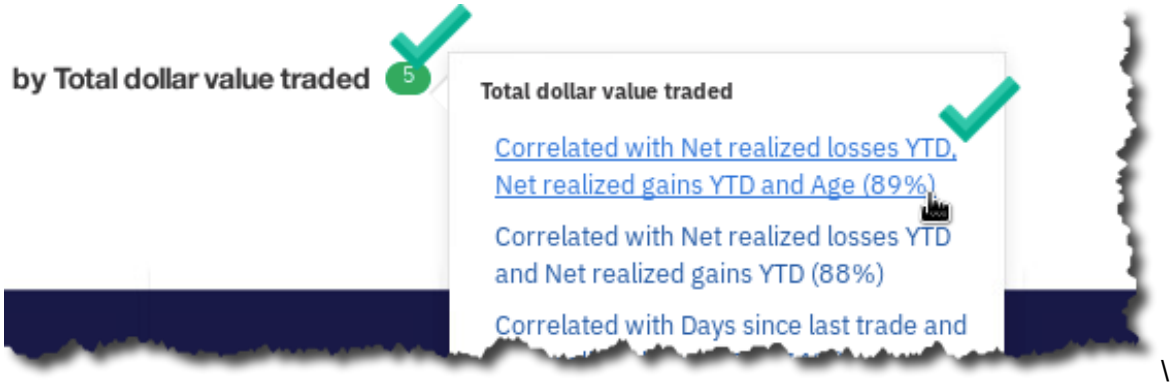
__41. From here, drag Gender over to on top of Status.



__42. This chart is showing that our high-risk customers who trade the most dollars with us are Female and Married. Cognos Analytics also provides hints at other data points which seem to be related. Cognos Analytics found 5 other points of interest for you. Select the `5` at the top and choose `Correlated with Net realized losses YTD, Net realized gains YTD and Age (89%).`

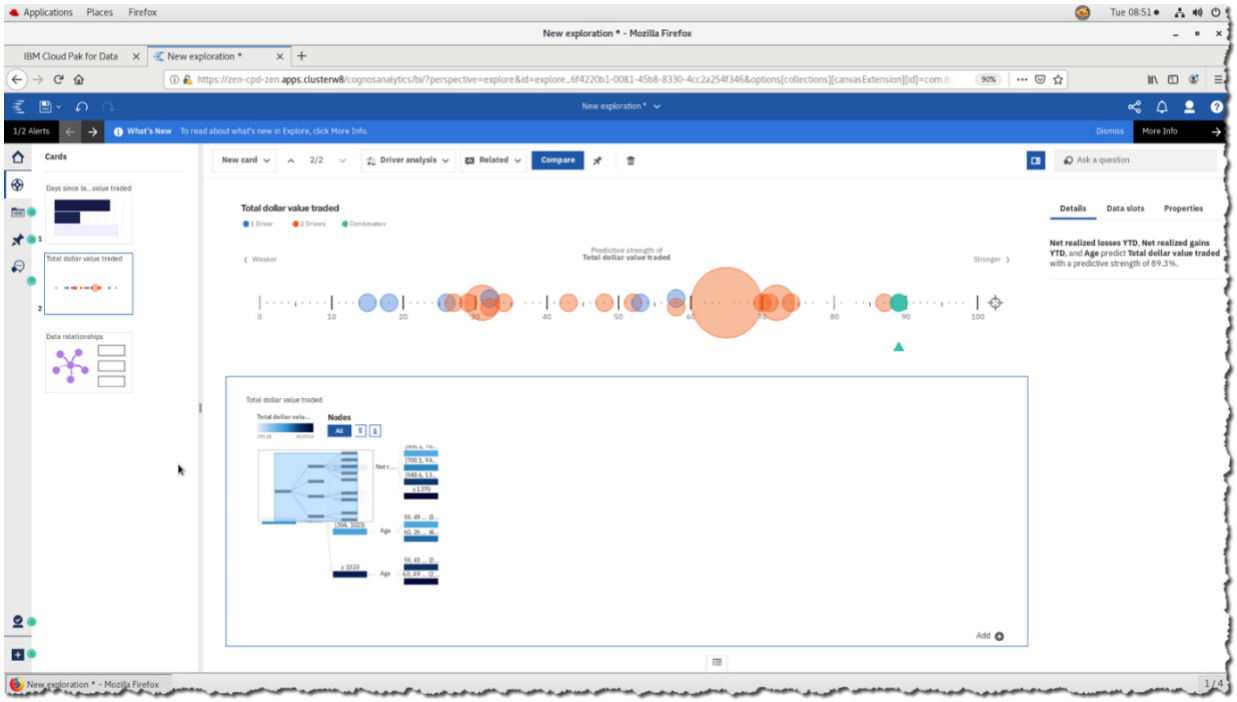


__43. Next to the title for the fly out, select the `5` ⇨ `Correlated with Net realized losses YTD, Net realized gains YTD and Age (89%).`



\

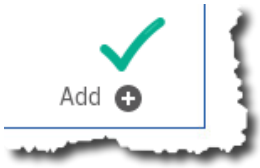
__44. Choose the `Explorations` tab on the left  and you will now see the following:



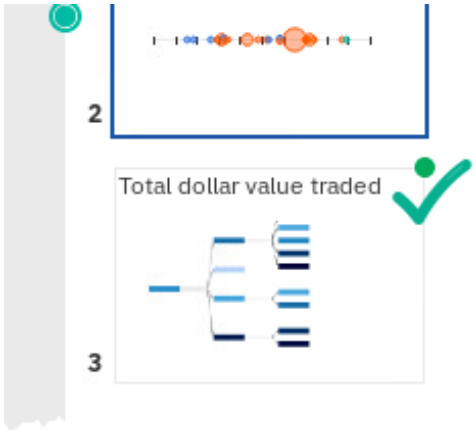
__45. Notice the strength of the predictor is higher when a combination of drivers is applied.

__46. Examine the Decision Tree at the bottom provided by IBM Cognos Analytics.

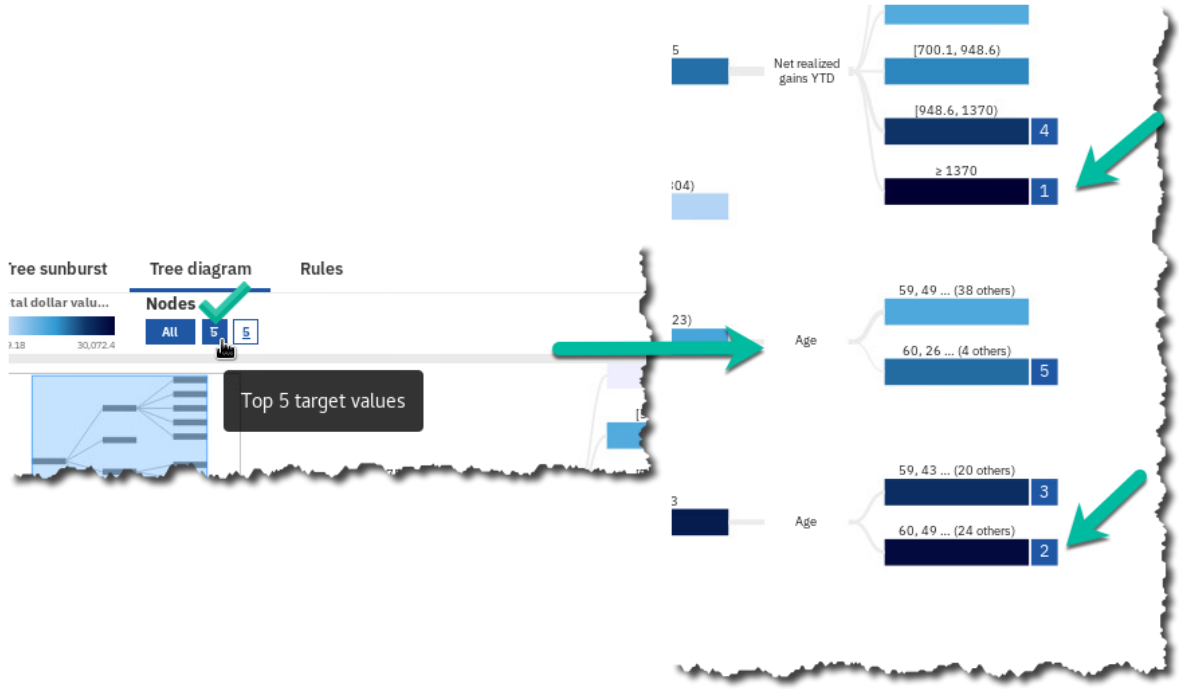
__47. At the bottom right, select `Add button` (to add this Decision Tree as a new Exploration card).



__48. Select the `Decision Tree` card at the left.

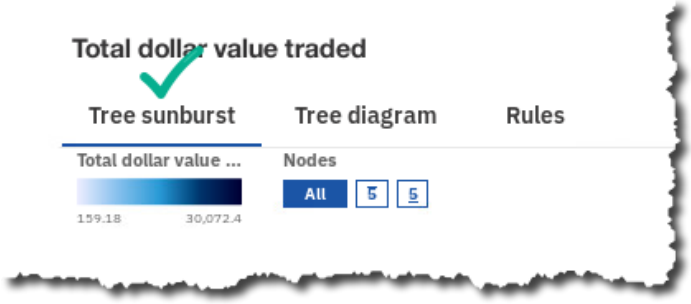


__49. Now you are easily seeing combinations of drivers and their relationship strength in terms of Total dollar value traded. Select the Top 5 Nodes to show your top 5 nodes in the tree.

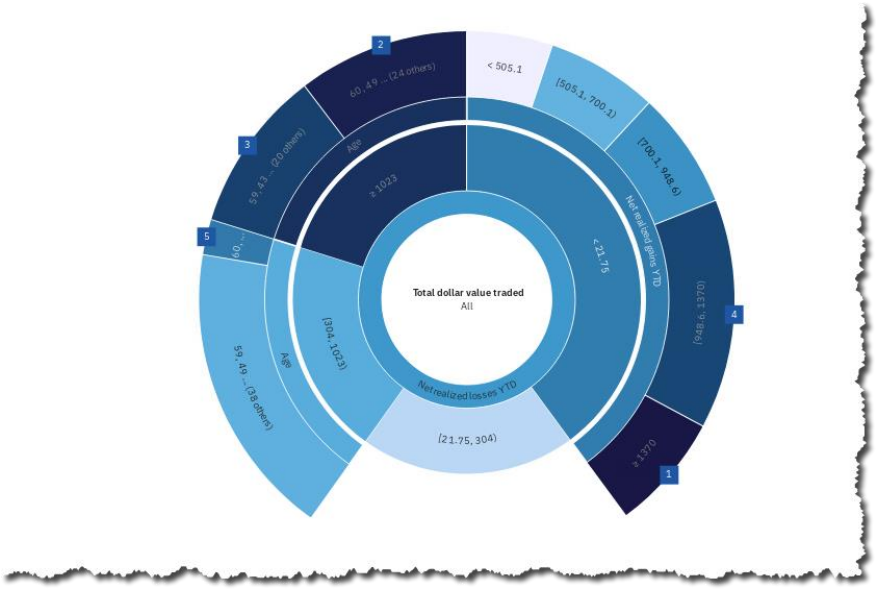


__50. But you are not limited there. You can view other ways of looking at this data.

Choose `Tree sunburst` from the top tab.

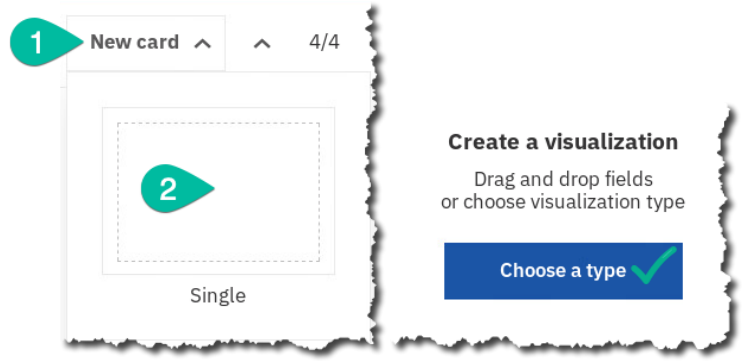


__51. You will now be presented with the Sunburst diagram.



__52. So far, IBM Cognos Analytics has been guiding us and creating charts based upon a metric we were looking for. We also have the ability to create exploration cards on our own.

__53. From the top menu, select `New card` ⇨ `Single – Choose a Type.`

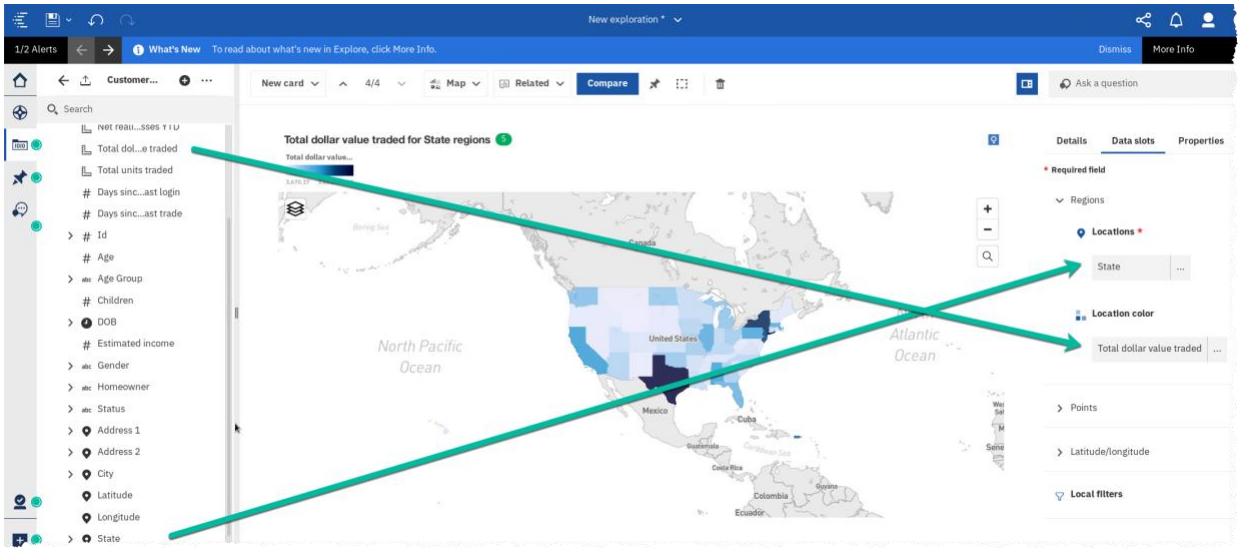


__54. Choose `Map` under the group `Geospatial.`

__55. Select the Data Tab  , then under Data slots, expand Regions.

Drag Total dollar value traded to Location color.

Drag State to Locations.



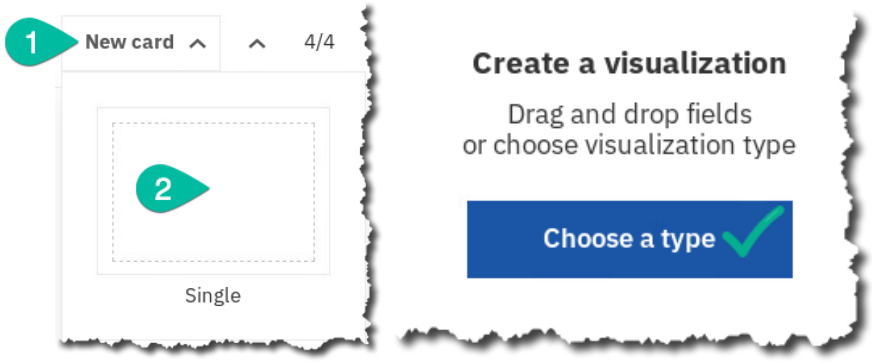
__56. Select the Details tab to get more information on this map.

It looks like most of our big spenders are in Texas and New York.

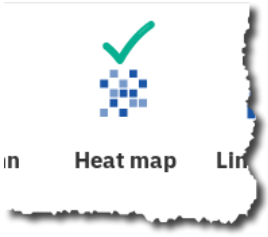
Zoom in on the U.S. and center it to focus.

__57. Let's create one more Exploration card.

From the top tab, select `New card` ⇨ `Single` ⇨ `Choose a type.`



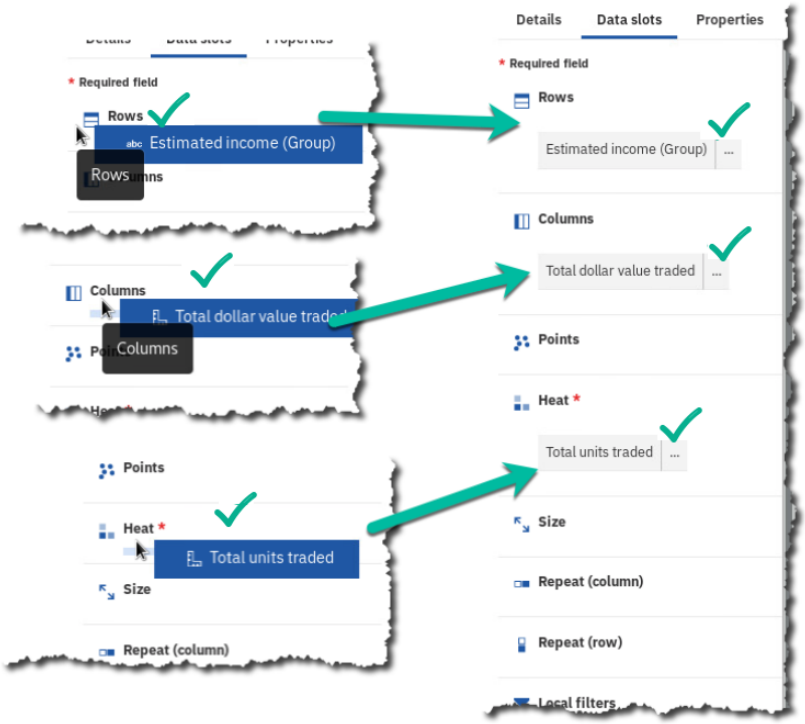
__58. This time let's choose a `heat map` from the `Relationships` group.



__59. Drag `Estimated income (Group)` to `Rows.`

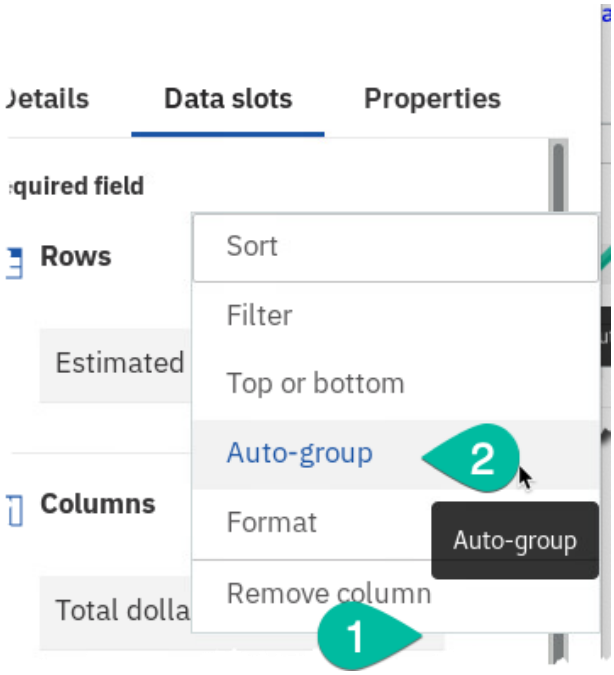
Drag `Total dollar value traded` to `Columns.`

Drag `Total units traded` to `Heat.`

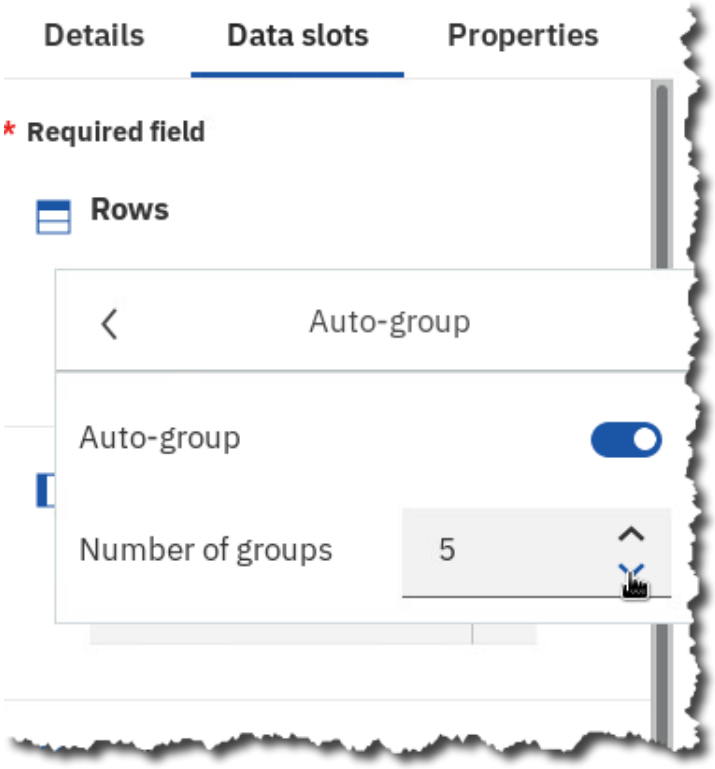


__60. That looks really noisy! Let's now group that `Total dollar value traded`. Lucky for us, IBM Cognos Analytics allows you to Auto-group on the fly.

Select the `ellipses` next to `Total dollar traded` ⇨ `Auto-group`.



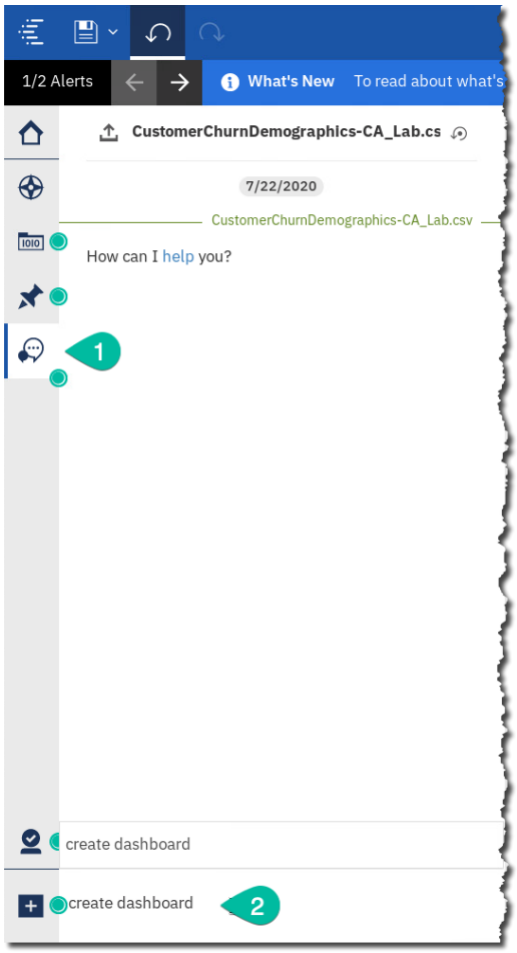
__61. Turn on `Auto-group` and set `Number of groups` to `5.`



__62. That's easier to read! Visualizing that sweet spot is now so much easier.

__63. Now we will have IBM Cognos Analytics Assistant create a Dashboard for us.

Select the `Assistant` link on the left, then choose the suggestion of `create dashboard.`

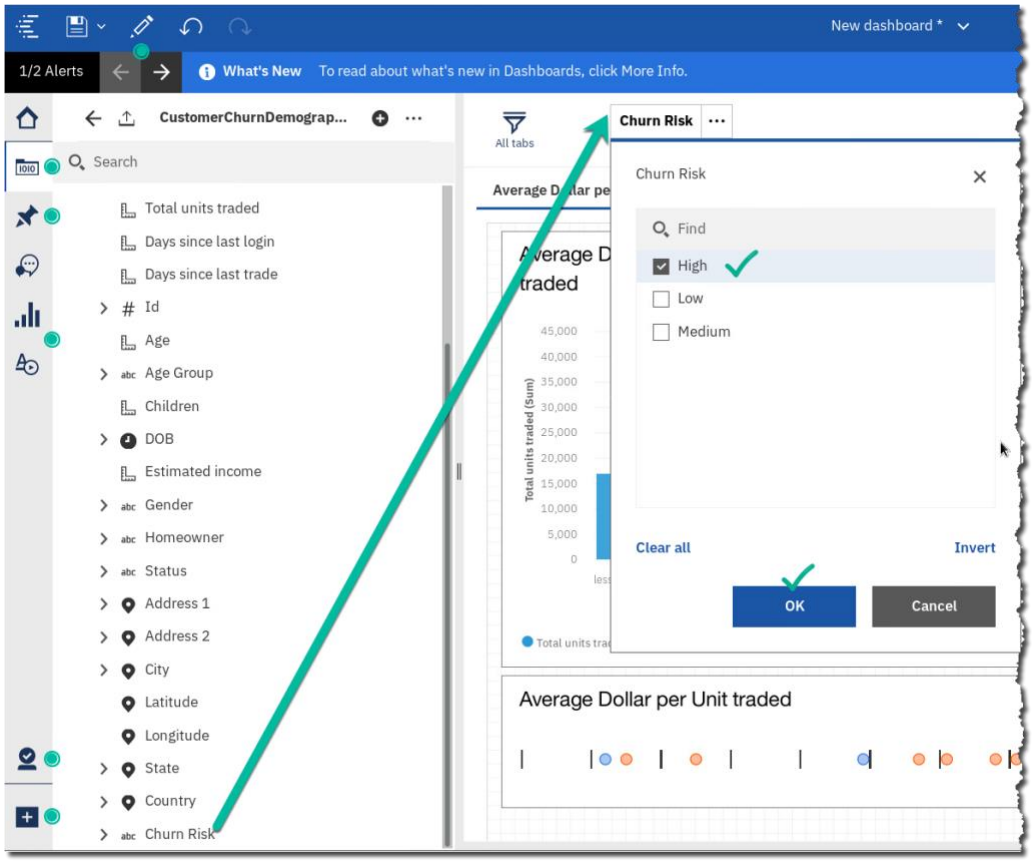


__64. You should now see the following:

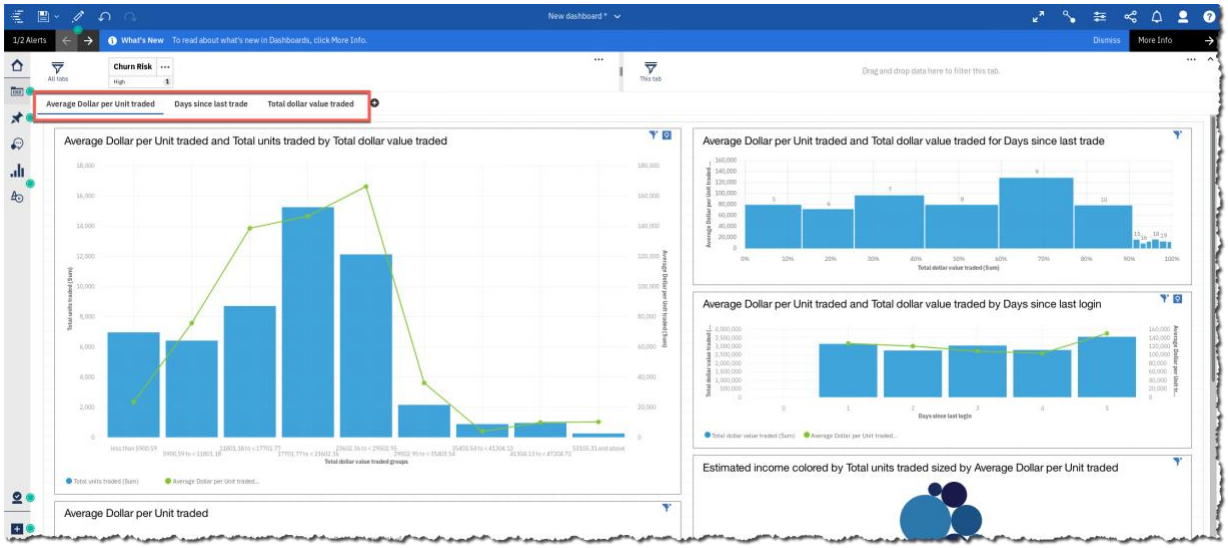
IBM Cognos Analytics Assistant created a dashboard for you!

__65. You can investigate this dashboard by selecting the tabs at the top or hover over any of the charts and graphs. We wish to focus in on our high churn customers. So, let's filter this dashboard.

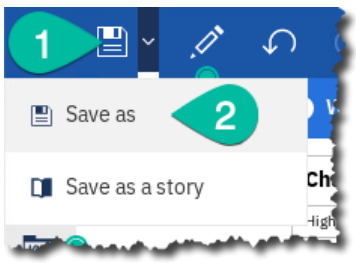
From the `Data` tab  , drag `Churn Risk` to the Filter `All tabs` location ⇨ `OK.`

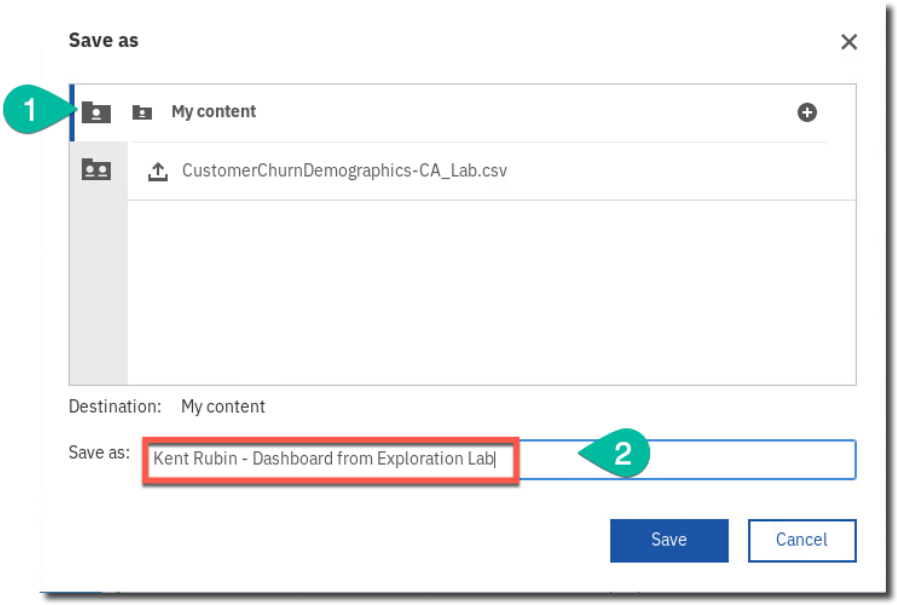


__66. Close the `Data` items tab by selecting it again and you should see all of your charts and tabs filtered on just your high churn risk customers. Building dashboards is easy in Cognos Analytics!



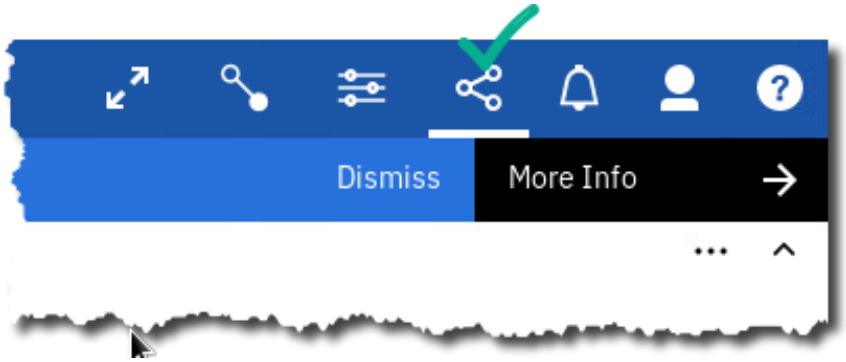
__67. `Save` your Auto-created Dashboard in your `Personal Folder` as `Your name – Dashboard from Exploration Lab.`





9.10 Embedding Cognos Content

__68. Now that you have created and saved your dashboard, you can Share your information to others. There are two options when sharing a link. You can either provide an HTML link to your dashboard or you can view the embedded code. You can also export your dashboard to a PDF file. To view all these options, choose the Share button at the top left.



__69. You will then see the Link options in the flyout

__70. Use this information to embed and link your Cognos Analytics dashboard.

9.12 Lab conclusion

The results of this Exploration and Dashboard give us a good understanding of our high churn risk customers, how much they spend and where they are located. I can make marketing decisions in these areas to ensure my business is not at risk of losing these customers.

You have now witnessed the ease at which IBM Cognos Analytics can create highly visual dashboards for greater business insight…all without need to code any expressions in a secure, governed manner.

You're off to a good start. Keep in mind, IBM Cognos Analytics offers insight and infused AI using several governed, easy to use components. Through this lab, you have experienced working with Data Modules, Explorations and Dashboards. But IBM Cognos Analytics offers other areas of insight as well, such as Reports and other areas of analysis that could not fit in this lab.

We encourage you to delve deeper into all these other avenues of insight by taking on other online IBM Cognos Analytics workshops available to you.

TIP: For a more detailed workshop on IBM Cognos Analytics Dashboards, please see:
http://ibm.biz/Cognos-Dashboards
TIP: For a more detailed workshop on IBM Cognos Analytics Reporting, please see:
http://ibm.biz/Cognos-Reporting

**** End of Lab 09 – Infuse: Cognos Analytics - Introduction**

Lab by Kent Rubin, IBM

Lab 10 WRAP-UP

10.1 Lab overview

Let's do some wrap up tasks the various personas might do after completing the entire set of "core" labs for this workshop.

Note: if you have not completed all the "core" labs in this workshop, doing some of these exercises may not make sense.

10.2 Data Scientist wrap-up

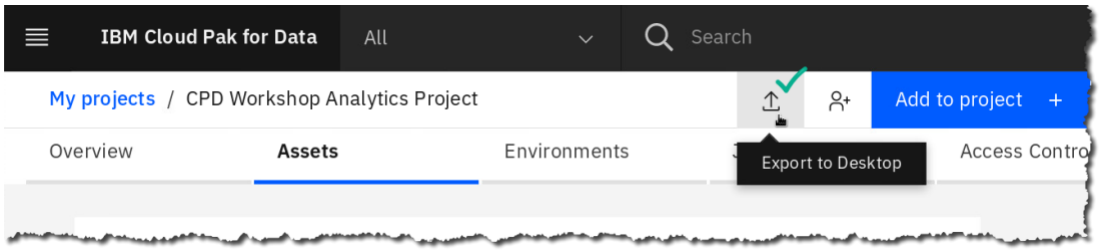
Data Scientist

10.2.1 Saving an analytics project

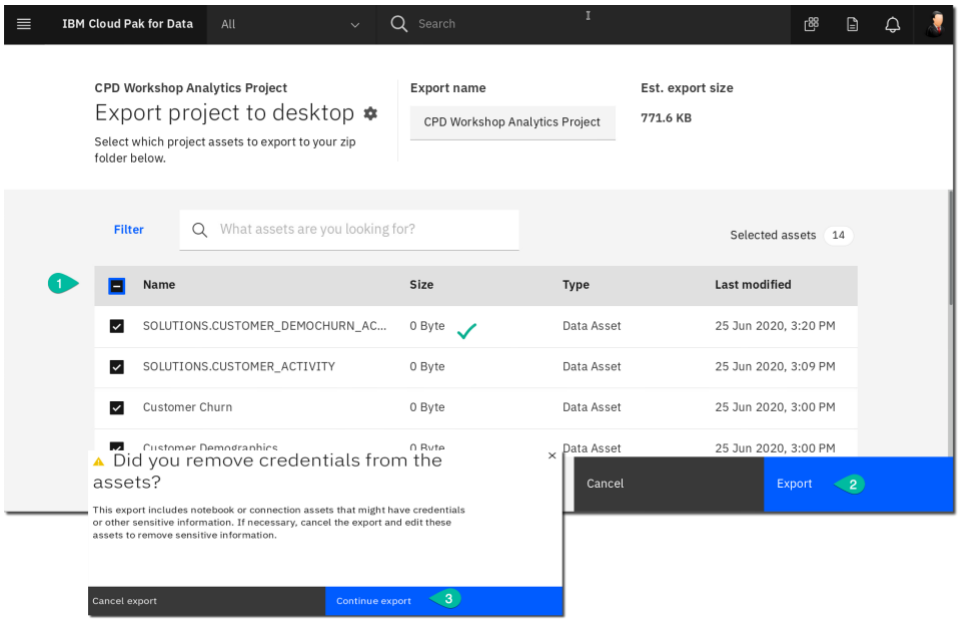
The Data Scientist saves her work by `Exporting` the project she just worked on to a file on the server in a location that the OS administrators consistently backup.

__1. Click `Navigation Menu` ⇨ `My Projects` ⇨ `CPD Workshop Analytics Project.`

Click on the `Export to Desktop` (up arrow) `icon` (at the top of the screen).

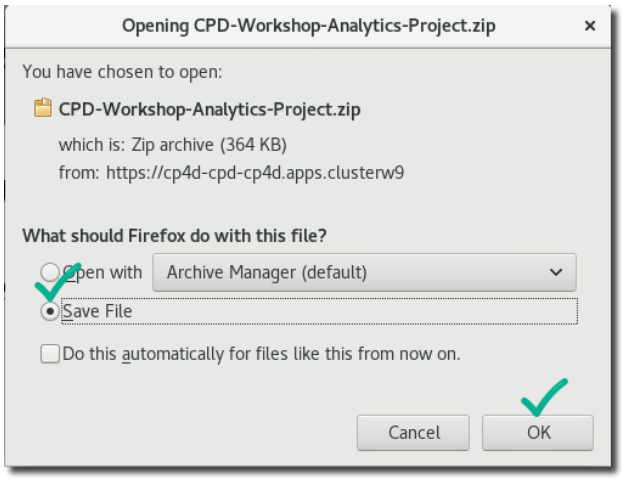


__2. Click the first box to select all the assets and then `Export` ⇨ `Continue export.`



__3. Click `Save file` ⇨ `OK.`

(Note: the file will be in `/home/ibmdemo/Download.`)



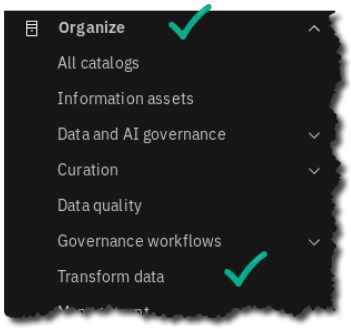
10.3 Data Engineer wrap-up

Data Engineer

10.3.1 Saving a transform project

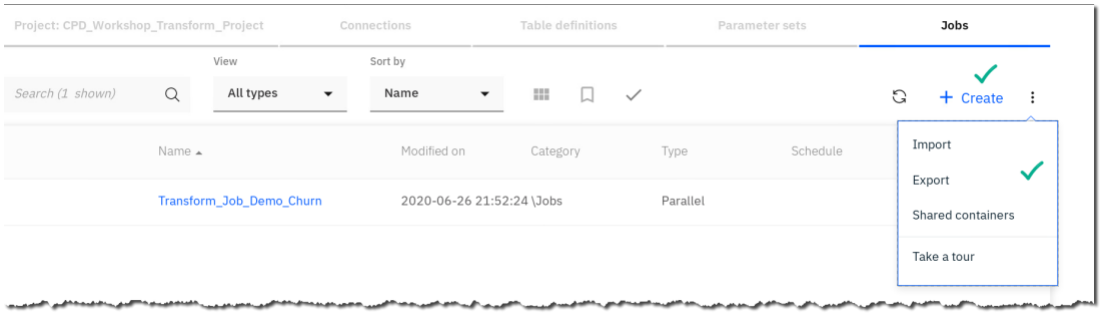
The Data Engineer saves her work by `Exporting` the project she just worked on to a file on the server in a location that the OS administrators consistently backup.

__4. Click `Navigation Menu` ⇨ `Organize` ⇨ `Transform data.`



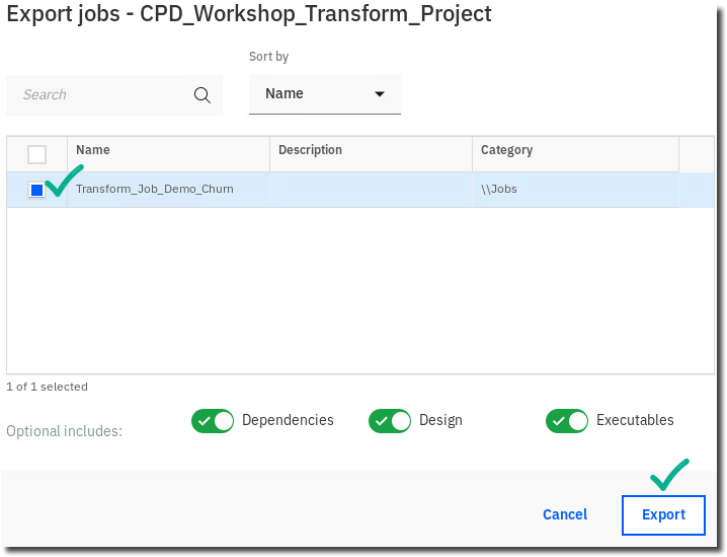
__5. Click on `CPD_Workshop_Transform_Project.`

__6. Click `ellipses` ⇨ `Export.`



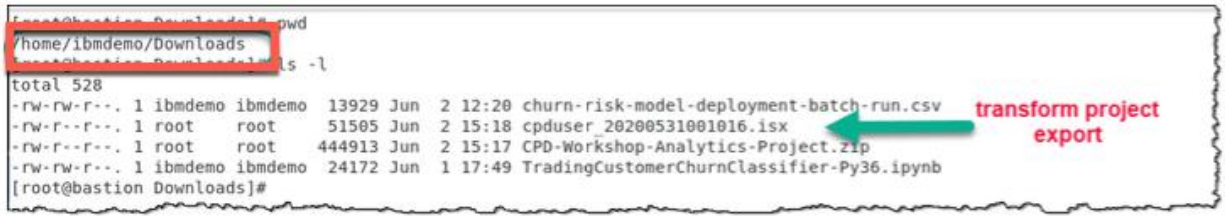
__7. Here is where you choose which jobs in the project to Export. You can click them all with the top checkbox. Since there is only one job in this project, checking either box works the same.

Click `Export.`

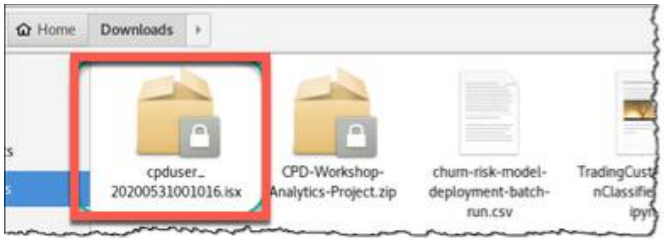


__8. The export is kicked off in the background. When it finishes, the file will be in `/home/ibmdemo/Downloads` if you are accessing via remote desktop. If using browser, you can save wherever you wish.

__9. Review it in a terminal window.



__10. You can also review it in your desktop browser.

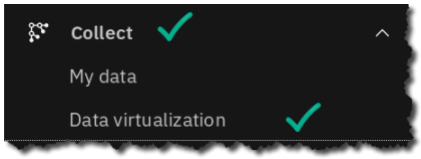


10.4 Administrator wrap-up

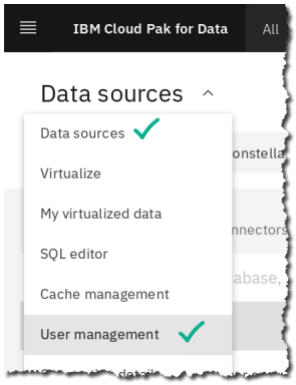
10.4.1 Granting Data Virtualization access

The Administrator has been asked to open up Data Virtualization capabilities to a trusted user called `Data Steward` which allows the user to create them himself.

__11. The `Navigation Menu` ⇨ `Collect` ⇨ `Data virtualization.`

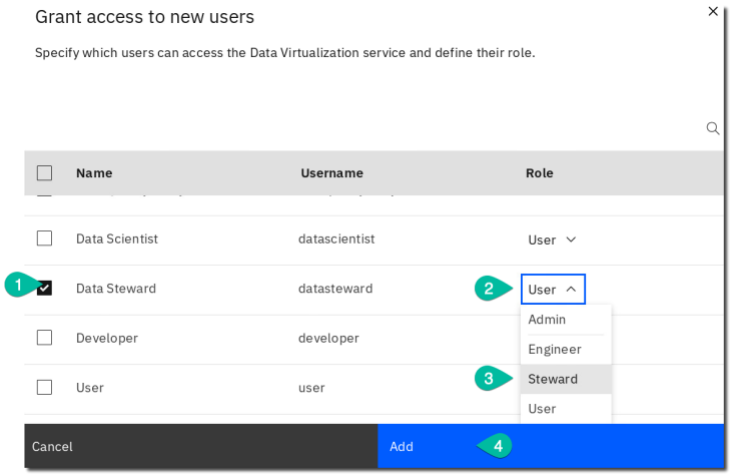


__12. Click the `Menu dropdown` (which now says: `Data sources`) ⇨ `User management.`



__13. Click `Add users +.`

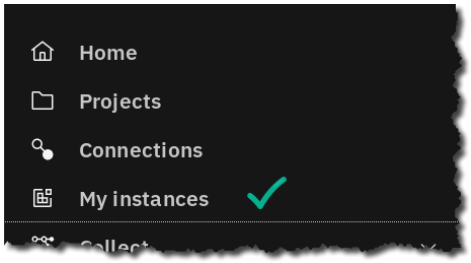
__14. Select Name: `Data Steward` ⇨ Role: `Steward` ⇨ `Add.`



10.4.2 Environment management

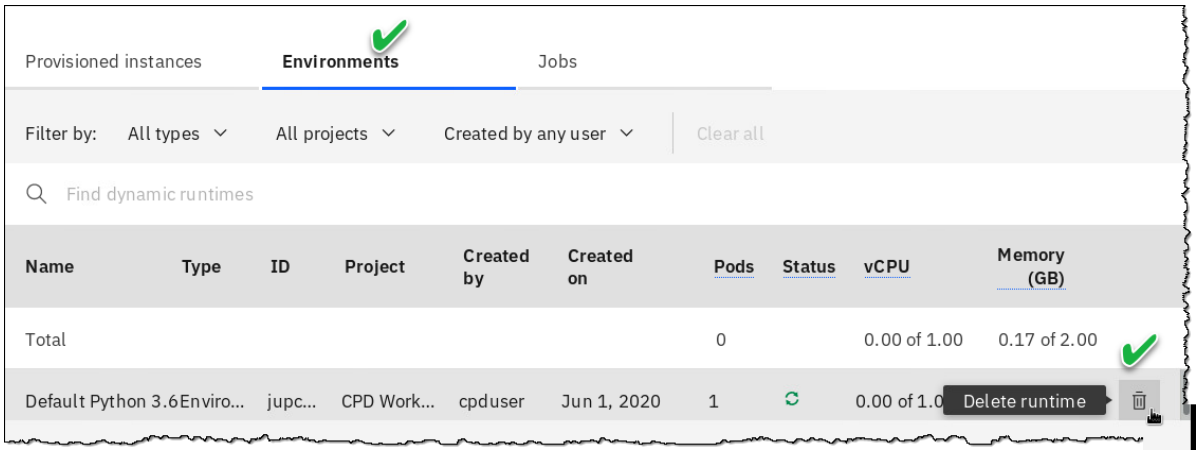
The Administrator may need to reclaim resources for environments that are long running and not in use.

__15. Click `Navigation Menu` ⇨ `My Instances.`



Click the Environments tab.

__16. Here you can delete any runtime environment that has been left running by a user to free up resources on the cluster.



10.5 Workshop conclusion

With this workshop, you can now see how Cloud Pak for Data turns your organization's data into a critical corporate asset with end-to-end data integration and collaboration on a modern, cloud-native platform.

You have now completed your IBM Journey to Cloud and AI: Analytics Modernization using Cloud Pak for Data.

**** End of Lab 10 - Wrap-up**

Lab by Burt Vialpando and Kent Rubin, IBM

Back Page: 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
Armonk, NY 10504-1785
USA

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106-0032, Japan

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 Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites 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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have

been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

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 of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental. All references to fictitious companies or individuals are used for illustration purposes only.

Back Page: Trademarks and Copyrights

The following terms are trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM	AIX	CICS	ClearCase	ClearQuest	Cloudscape
Cube Views	Db2	developerWorks	DRDA	IMS	IMS/ESA
Informix	Lotus	Lotus Workflow	MQSeries	OmniFind	
Rational	Redbooks	Red Brick	RequisitePro	System i	
System z	Tivoli	WebSphere	Workplace	System p	

Adobe, Acrobat, Portable Document Format (PDF), and PostScript are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, other countries, or both.

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.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both. See Java Guidelines

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

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.

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

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

ITIL is a registered trademark and a registered community trademark of the Office of Government Commerce and is registered in the U.S. Patent and Trademark Office.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Other company, product and service names may be trademarks or service marks of others.

NOTES

NOTES

Please Recycle

IBM Software