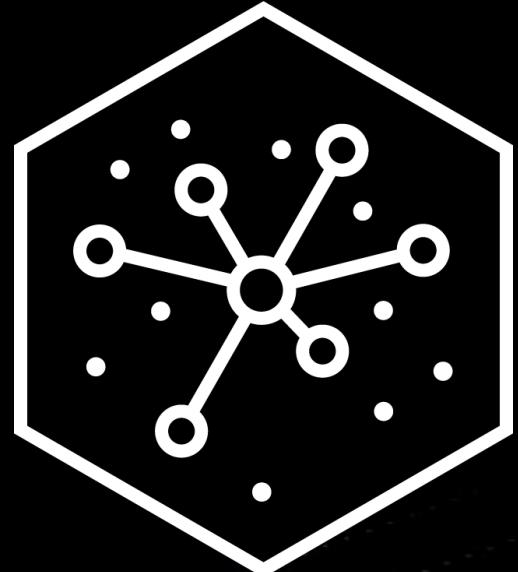


# IBM Journey to Cloud and AI Analytics Modernization Workshop

Featuring: Cloud Pak for Data 3.0.1



Session starts at 9 AM

# Welcome to the IBM® Briefing Center

## Location logistics

- ✓ Access restrictions
- ✓ Restrooms
- ✓ Emergency exits
- ✓ Smoking policy
- ✓ Breakfast / Lunch / Snacks
- ✓ Special meal requirements

## Introductions

- ✓ IBM Speakers
- ✓ IBM Proctors
- ✓ IBM Sales Reps
- ✓ Attendees (optional)

# IBM Analytics Modernization Workshop

## Agenda

<b>Part 1</b>	<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
<b>Part 2</b>	<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deep Dive</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
<b>Part 3</b>	<ul style="list-style-type: none"><li>• Analyze</li><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 06</li><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

# Zoom tips for today's session

## The Zoom Tool Bar



We will stay on mute during the main sessions. You may unmute yourself during breakouts.

We will not use video to preserve bandwidth during our class today.



Participants (2)

A screenshot of the Zoom participants list. It shows two entries: 'ciarrocc@us.ibm.com (me)' with a green profile picture and 'Jennifer Ciarrocca (Host)' with a green profile picture. To the right of each entry are two small red icons: a crossed-out microphone and a crossed-out video camera.

Chat

From Me to Everyone:

Here is a test chat that everyone will be able to see. This is where you can type a question during class

A screenshot of the Zoom chat window. It displays a message from the host: 'From Me to Everyone: Here is a test chat that everyone will be able to see. This is where you can type a question during class'. The text is in a black font on a white background.

The Share Screen function has been deactivated during the lectures

# IBM Analytics Modernization Workshop

## The workshop Unified Desktop

The screenshot shows a dark-themed desktop environment with various application icons and a central window for the 'IBM Journey to Cloud and AI Analytics Modernization' workshop.

**Central Window:**

- Title:** IBM Journey to Cloud and AI Analytics Modernization
- Version:** Workshop v9.0
- Subtext:** Featuring: Cloud Pak for Data

**Available Applications:**

- IBM Cloud Pak for Data
- Watson Assistant-Discovery
- OpenShift Web Console
- Terminal
- Lab Solutions
- Home

**Installed Software:**

- Cloud Pak for Data v3.0.1 Enterprise
- OpenShift (RHOCP) v3.11.219
- Kubernetes v1.11
- Red Hat Enterprise Linux Server Release 7.8 (Maipo)
- Netezza Performance Server 11.0.3.1

**Installed IBM Cloud Pak for Data Services:**

- Watson Knowledge Catalog
- Decision Optimization
- Watson Studio Local
- Db2 Advanced Edition
- Watson Machine Learning
- MongoDB
- Watson OpenScale
- Cognos Analytics
- Data Virtualization
- Cognos Dashboard Embedded
- DataStage Edition
- SPSS Modeler
- Metrics Server

**Icons on the desktop:**

- Cluster Configuration.png
- HCCX-cluster-xx
- Trash

**Bottom Left:** IBM logo

### Workshop lab exercise tips

- Use full screen with a large screen computer
- Use Chrome or Firefox browsers
- Use [Ctrl][+]/[Ctrl][-] or [Ctrl][Mouse-Scroll-Wheel] to zoom
- Use a mouse (keyboard alone is difficult)

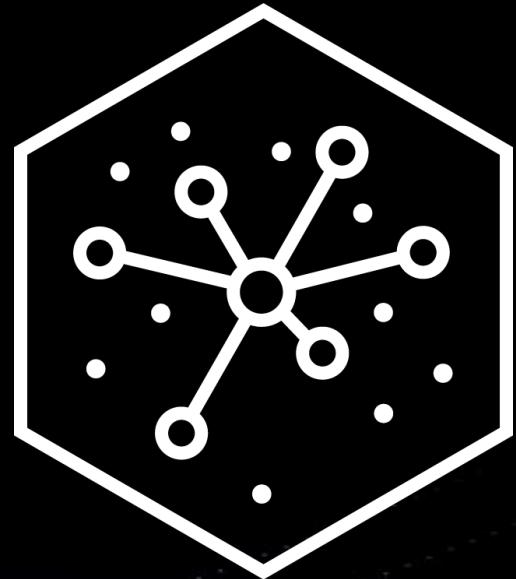
# IBM Analytics Modernization Workshop

## Part 1

- |   |  |
|---|--|
| <ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>  | <ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>  |
| <ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deeper Dive</li><li>• Collect: Virtualize</li></ul>                             | <ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>                                   |
| <ul style="list-style-type: none"><li>• Analyze</li><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul> | <ul style="list-style-type: none"><li>• Lab 06</li><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul> |

# Introduction

*Lab 01 – Getting Started*



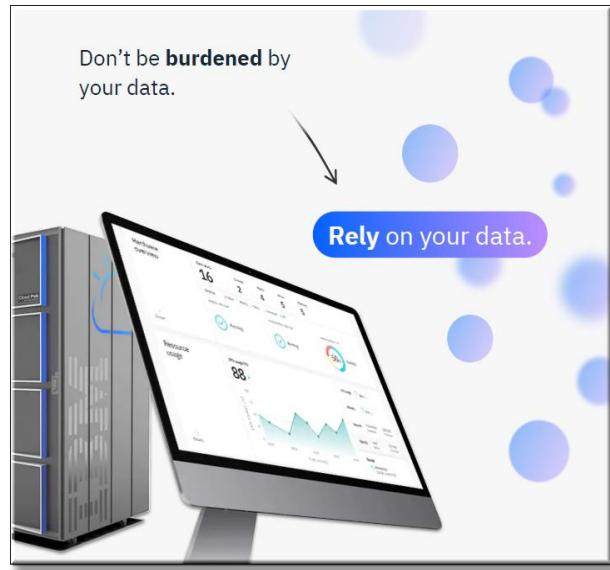
# IBM Cloud Pak for Data



**IBM Cloud Pak for Data** is a single unified, integrated platform which helps to simplify the collection, organization and analysis of data.

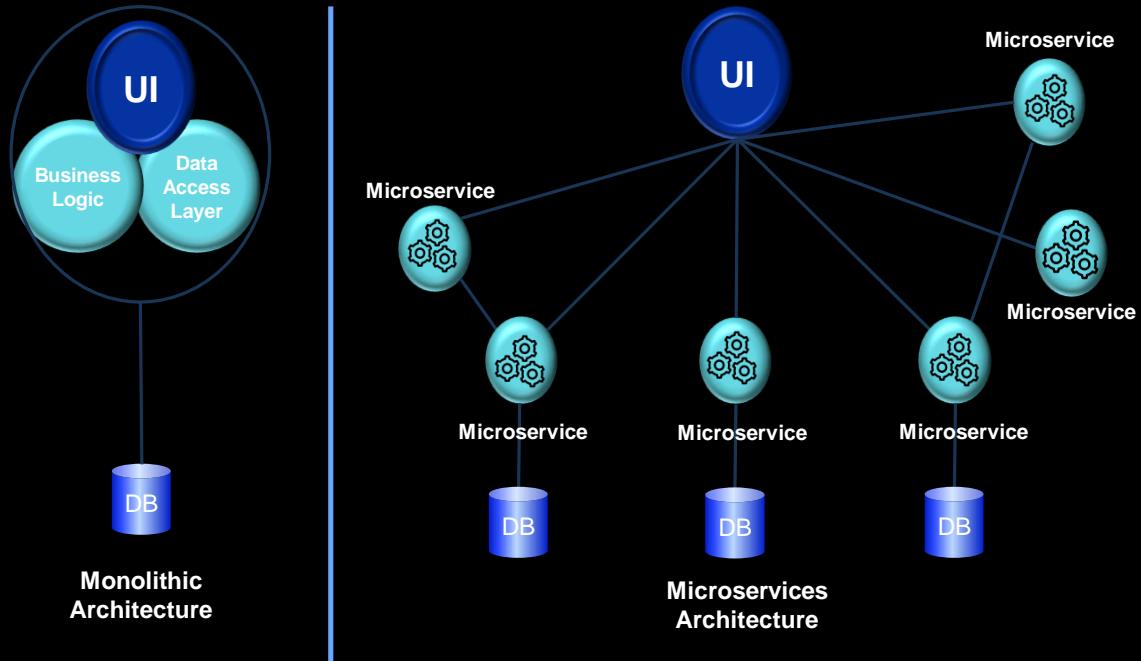
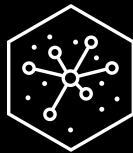
With it, enterprises can turn data into insights through an integrated cloud-native architecture.

IBM Cloud Pak for Data is extensible and easily customized to unique client data and AI landscapes through an integrated catalog of IBM, open source, and third-party microservices.



# Microservices – the first key to cloud native applications

## Making development & deployment more efficient



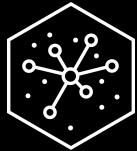
### Microservices benefits \*

- Improved fault isolation:**  
Larger applications can remain largely unaffected by the failure of a single module
- Technological flexibility:**  
Try out a new technology stack on an individual service and roll it back if required
- Easier development:**  
A new developer can more easily understand the functionality of a service
- Optimized deployment:**  
Auto provision, auto scale and provide auto-redundancy

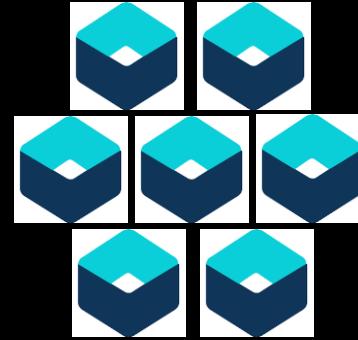
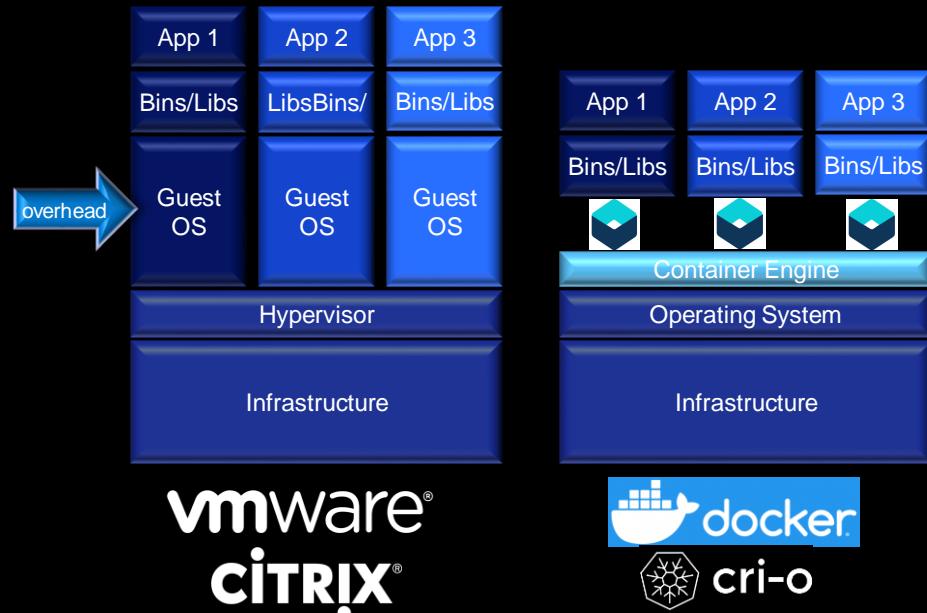
\* This is not a claim that a microservice-based application approach is always better for every use case scenario

# Containers – the second key to cloud native applications

## Reducing operational and development costs



### Virtual machines vs. containers \*



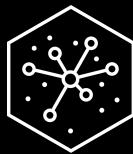
Containers can be 2 – 3 times more resource efficient than virtual machines

On average Docker developers ship software 7x more frequently

\* Containers virtual software in the way that virtual machines have virtualized hardware

# Container automation and orchestration is essential

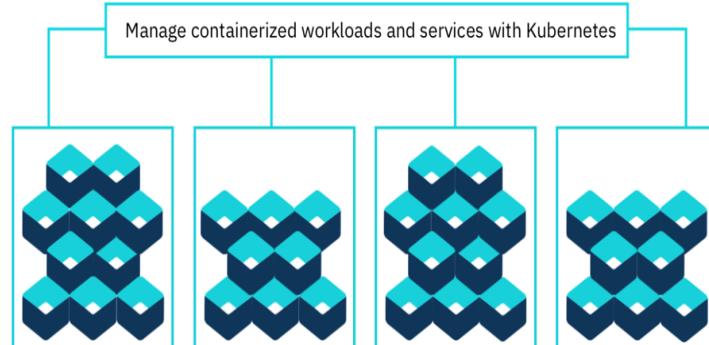
Enter: Kubernetes



**Containers are revolutionizing IT  
But they require orchestration**

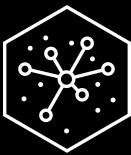


**Kubernetes - κυβερνήτης  
Means “helmsman” or “pilot”**



# Red Hat OpenShift

## Enterprise Kubernetes Platform



OPENSIFT

**Advanced Cluster Manager**

**OpenShift Container Platform**

**OpenShift Kubernetes Engine**

### Multi-cluster Management

Discovery : Policy : Compliance : Configuration : Workloads

Manage Workloads

Build Cloud-Native Apps

Developer Productivity

#### Platform Services

Service Mesh : Serverless Builds : CI/CD Pipelines Full Stack Logging Chargeback

#### Application Services

Databases : Languages Runtimes : Integration Business Automation 100+ ISV Services

#### Developer Services

Developer CLI : VS Code extensions : IDE Plugins Code Ready Workspaces CodeReady Containers

#### Cluster Services

Automated Ops : Over-The-Air Updates : Monitoring : Registry : Networking : Router : KubeVirt : OLM : Helm

#### Kubernetes

#### Red Hat Enterprise Linux & RHEL CoreOS



Physical



Virtual



Private cloud



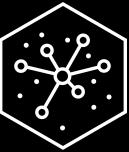
Public cloud



Managed cloud  
(Azure, AWS, IBM, Red Hat)

# Cloud Pak for Data (CPD)

## Make your data ready for AI



There is  
no **AI**  
without **IA**



**Infuse** - Deploy trusted AI-driven business processes



**Analyze** - Scale insights with ML everywhere



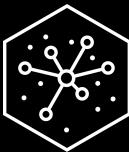
**Organize** - Create a trusted analytics foundation



**Collect** - Make data simple & accessible

**Strong Foundation – Built on “Cloud native architecture”**

# CPD Services Ecosystem



## 1. Cloud Pak for Data Base Services

Collect		Organize		Analyze		Deploy / Infuse	
Data Virtualization		Watson Knowledge Catalog (With IA, IGC, Refinery, InstaScan)		Watson Studio	Analytics Engine	Data Science: Model Design & Deployment	
Db2 Warehouse	PostgreSQL			Dashboards	IBM Streams		Watson OpenScale
Db2 Event Store	IBM Streams	Open Source Management		Industry Accelerators (many)			
Db2 Big SQL	NPS			Watson Machine Learning			
OpenShift / Control Plane (Lite)							

## 2. Premium Services (purchase license or BYOL)

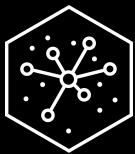
Collect	Organize	Analyze	Deploy / Infuse
Db2 AESE Virtual Data Pipeline	Infosphere DataStage Edition Infosphere Regulatory Accelerator Infosphere multi-cloud Data Mvmt Infosphere Entity Resolution Master Data Management	Cognos Analytics SPSS Modeler Decision Optimization Watson Explorer Planning Analytics	Watson Assistant / Discovery Watson API Kit (Speech to Text, Text to Speech, Natural Language Understanding) Watson Financial Crimes Insights Planning Analytics

## 3. Third Party Extension Services

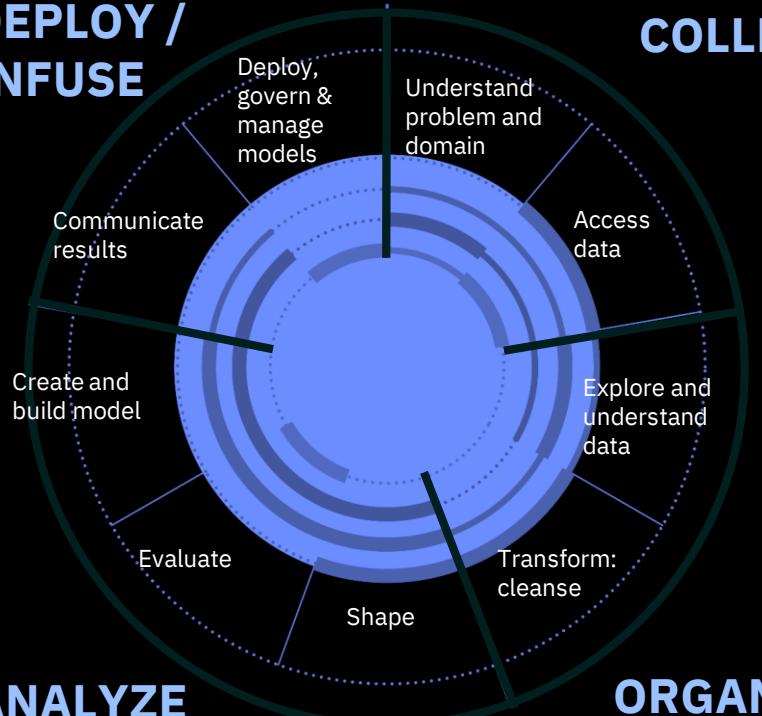


# Cloud Pak for Data (CPD)

Increases workforce productivity across the analytics lifecycle



## DEPLOY / INFUSE



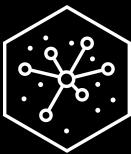
## COLLECT

## ANALYZE

Administrator / Architect	Data Engineer	Data Steward
Ensures the usability of the compute, network, storage, etc.	Architects data pipelines & ensures operability	Governs data and ensures regulatory compliance
Business Analyst	Data Scientist	Application Developer
Works with data to apply insights to business strategy	Dives deep into the data to draw insights for the business	Plugs into analysis and code to build applications

# CPD Administration

## Administer and manage the platform



The screenshot illustrates the CPD Administration interface, showing various management functions:

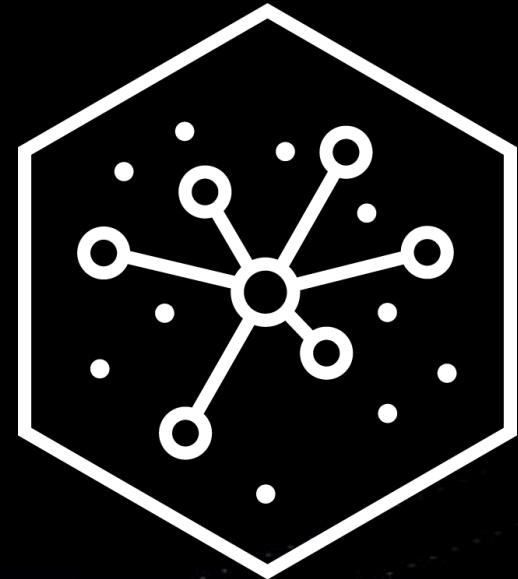
- Left sidebar (Administer section):**
  - Manage platform** (highlighted with a green oval)
  - Configure platform
  - Gather diagnostics
  - Manage users** (highlighted with a green oval)
  - Customize branding
- Deployments view:** Shows a list of deployments with columns: Name, Type, Installed on, Service instances, vCPU, and Memory (GB). It includes a search bar and filters for All types and Clear all.
- User Management view:** Shows a list of users with columns: Name, User ID, and Username.
- Pods view:** Shows a list of pods with columns: Name, Pods, vCPU, and Memory (GB).

Name	User ID	Username
admin	1000330999	admin
Business Analyst	1000331009	businessanalyst
CPD User	1000331002	cpduser
Data Engineer	1000331003	dataengineer

Name	Pods	vCPU	Memory (GB)
Total	8	0.14 of 1.65	1.50 of 2.80
db2wh-1590588027600-ibm-unified-console-api	5	0.10 of 1.00	1.30 of 2.00
db2wh-1590588027600-ibm-unified-console-influxdb	1	0.00 of 0.10	0.08 of 0.25
db2wh-1590588027600-ibm-unified-console-ucgoapi	1	0.00 of 0.25	0.07 of 0.10

# Business Use Case

*Lab 02 – Business Use Case: Customer Churn*



# Trade Co. Challenges

Customer retention problem leading to declining revenue

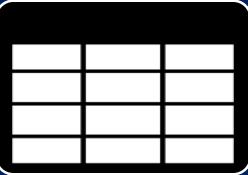
Underperforming rules-based system to identify separation (churn) risk

Lack of centralized, vetted, and reliable data to ensure accuracy of analytics

Disparate analytical tools for reporting and model development

No simple way to infuse machine learning models into the customer facing Stock Trader Application

# Separation (Churn) Risk: Current Rules Based System



## *Built Using Limited Data*

Rules are developed using a single source of data that contains customer demographic information.



## *Manual Process to Develop Rules*

Rules are manually developed based on the past experience of the marketing team. Rules are only updated once a year.



## *Low Overall Predictive Accuracy*

Low overall predictive accuracy. We are both missing identifying customers who ultimately separate and incorrectly assigning high risk to customers who ultimately stay.

# 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*

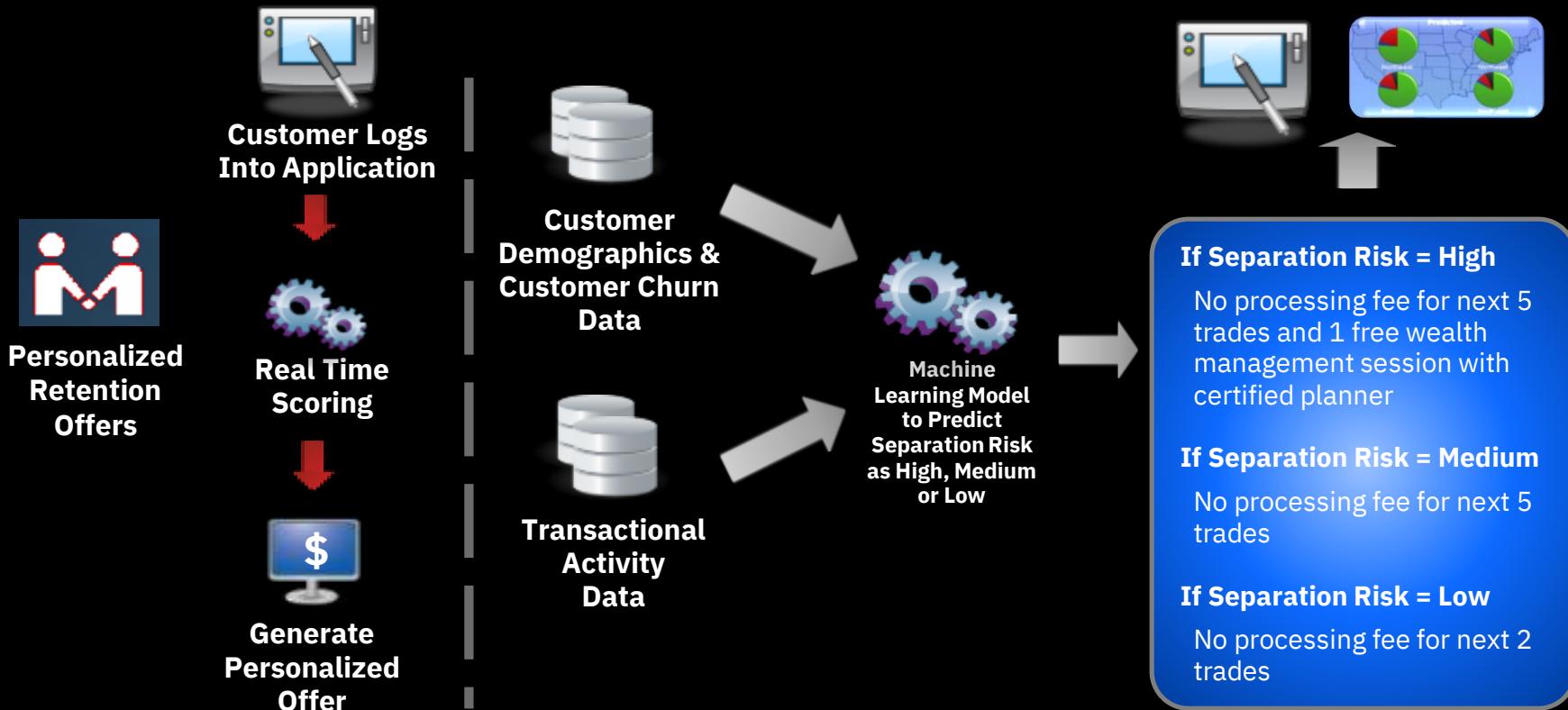
Develop predictive models for separation risk that automatically discover and incorporate all the patterns in the data including interactions and contingencies.



## *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.

# Deployment: Stock Trader App. Integrated with AI





Stock Trader Application

Stock Trader Application with Infused ML



Flat File of Monthly Sales Performance



1) Dashboard of Sales Performance (Monthly Metrics)



Customer Demographics

Historic Churn Risk Results



Customer Activity



2) Organize Data:  
Discover, Govern and Catalog



3) Transform Data:  
Merge and Prepare data for Analysis



4) Dashboard of Churn Risk (Demographics Discovery)



5) Build Machine Learning Model for Churn Risk (AutoAI and Notebook)

7) Integrate Model into Application (Stock Trader)



8) Dashboard of Business Impact (Monthly Metrics after AI)



Combined Data:  
• Demographics  
• Churn Risk  
• Activity



Post analysis customer data



**COLLECT**

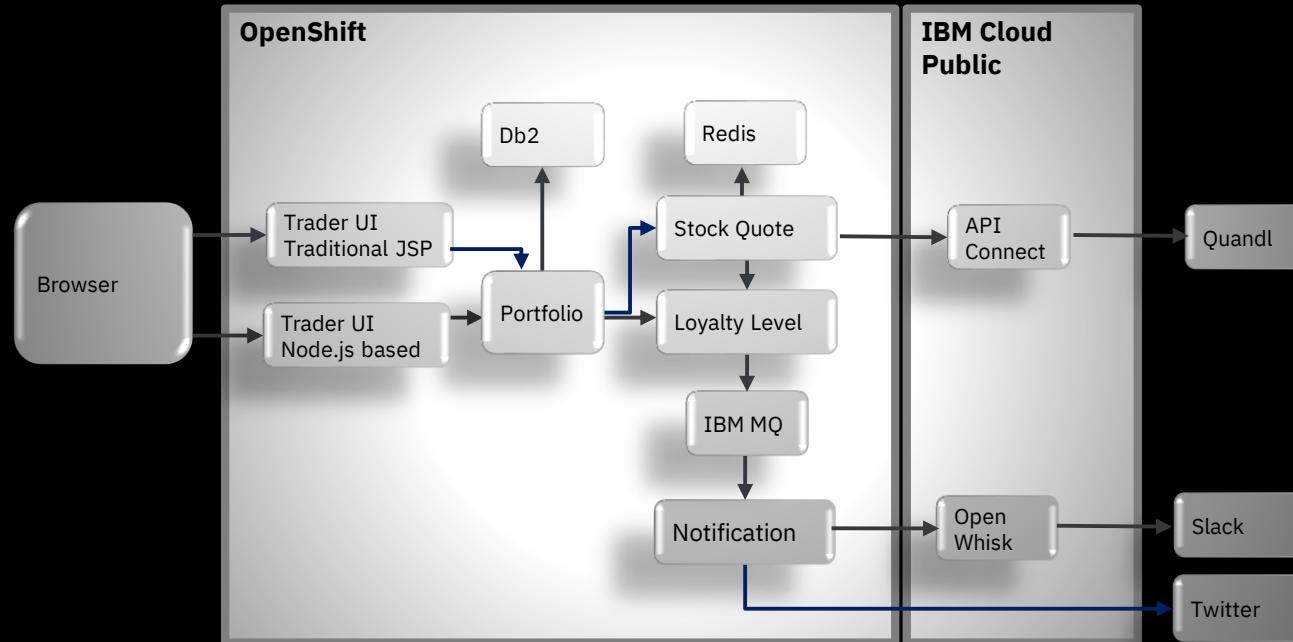
**ORGANIZE**

**ANALYZE**

**DEPLOY / INFUSE**

# Trade Co. Application “Stock Trader” – Before Microservices Application

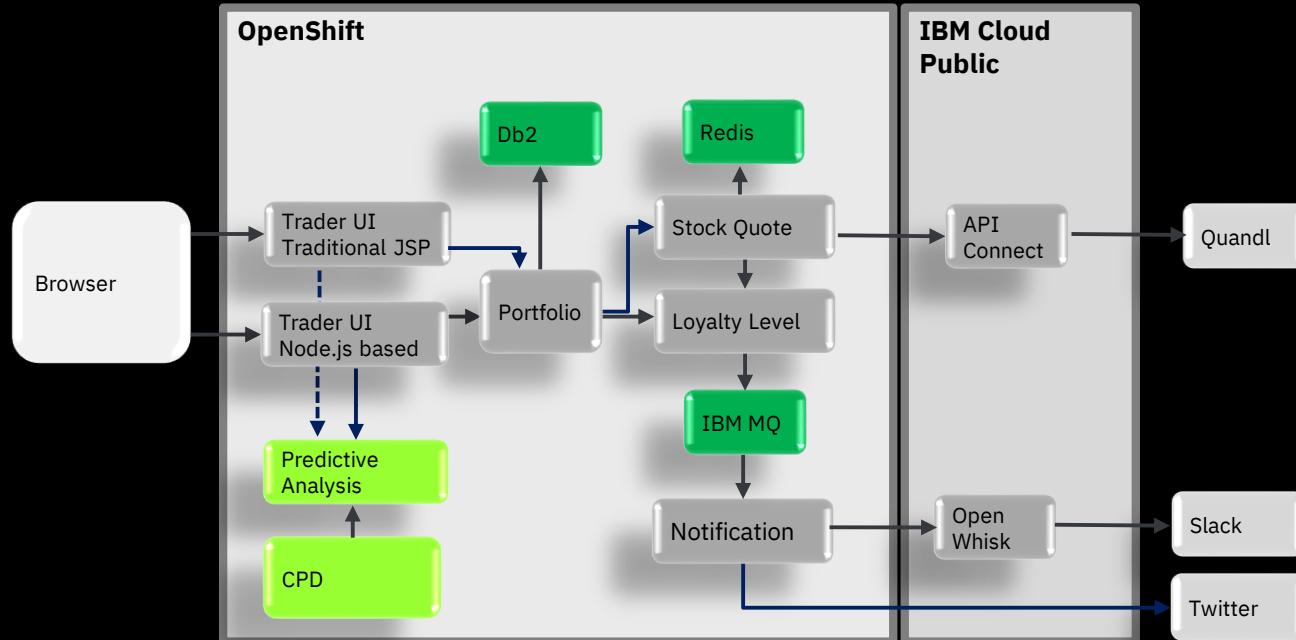
## Stock Trader as a modern microservice application



# Trade Co. Application “Stock Trader” – After: Infused with ML

## Microservices Application

### Stock Trader (enhanced) as a modern microservice application



# Stock Trader – After Monetizing the ML model

**IBM TRADER**

Home   Summary   Add Portfolio   Predictive Analysis   Change User

## Summary

Welcome to IBM Trader powered by ICP for Data

Create a new portfolio  
 Retrieve selected portfolio  
 Update selected portfolio (add stock)  
 Delete selected portfolio

Owner	Total	Loyalty Level
TechStocks	\$115,670	Gold

**Submit   Change User**

Though looking simple - a lot has gone through to provide machine learning predictive model scoring service.

no processing fee for next 5 trades

Advertisement

**IBM Cloud Pak for Data**

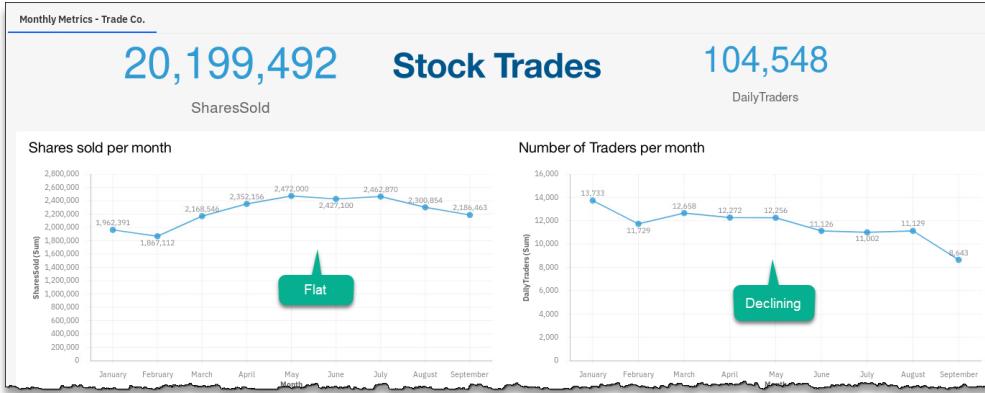
- Cloud agile
- Lightning fast
- AI-ready

No assembly required

# Trade Co. Dashboards

Before and After deploying the CPD developed ML model

Before AI

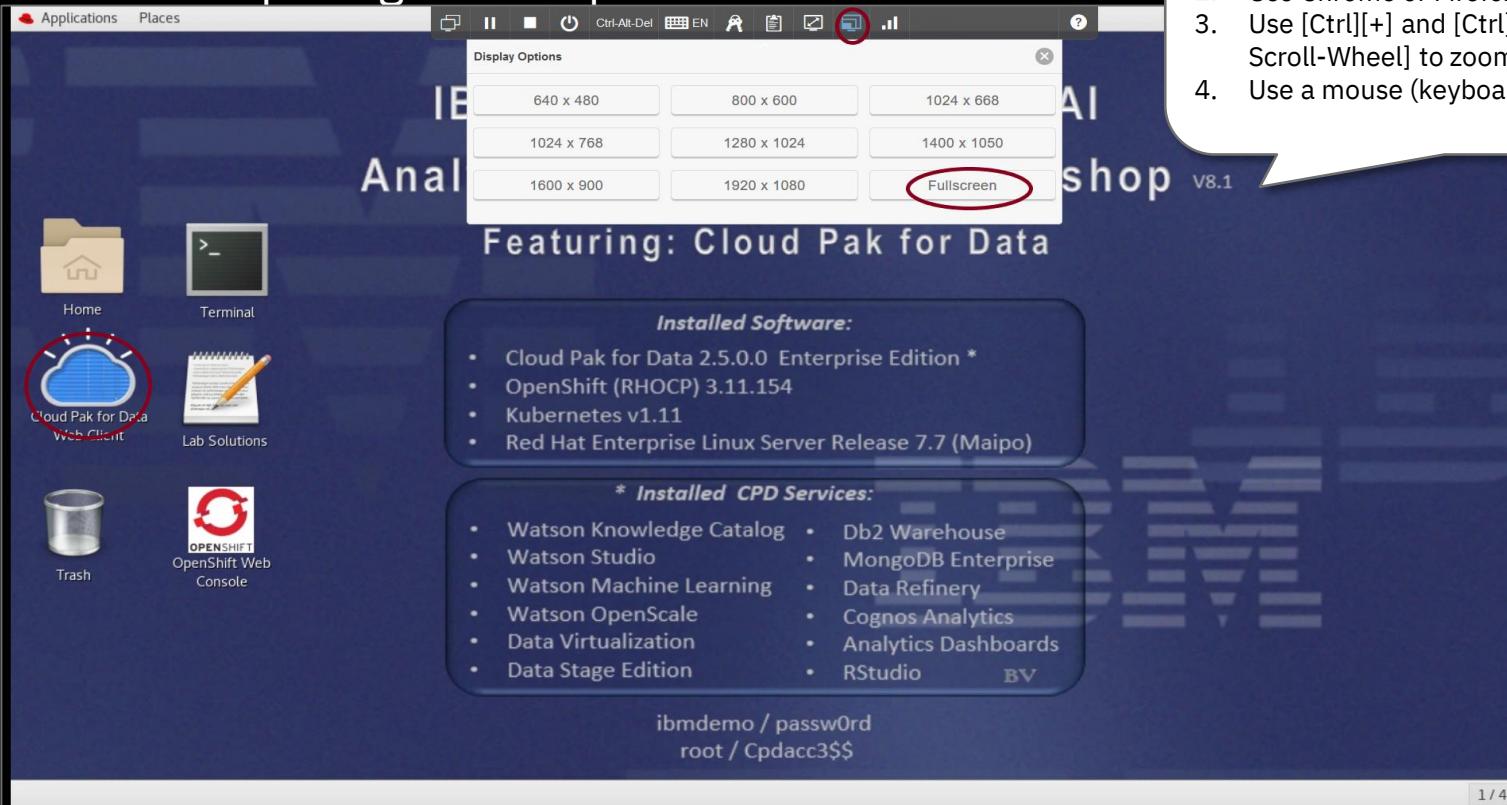


After AI

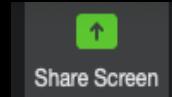
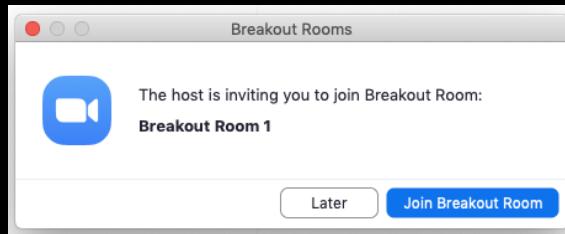


# IBM Analytics Modernization Workshop

## The workshop image desktop



## Zoom Breakout Rooms



Leave Breakout Room



Lab-01: Getting Started  
Case

Lab-02: Business



# IBM Analytics Modernization Workshop

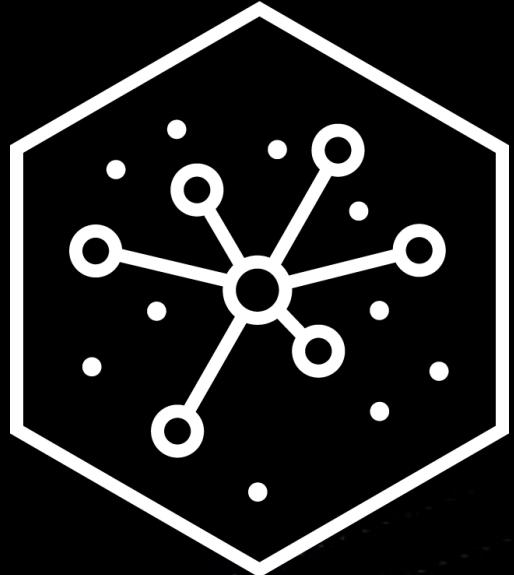
## Part 1

	<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
	<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
	<ul style="list-style-type: none"><li>• Analyze</li><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 06</li><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

# Cloud Pak for Data 3.0.1

Work on Lab 1 and 2

Session restarts at xx AM



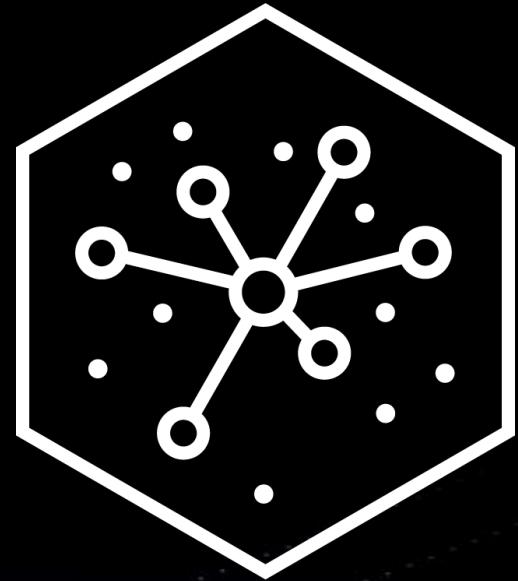
# IBM Analytics Modernization Workshop

## Part 2

	<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
	<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deep Dive</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
	<ul style="list-style-type: none"><li>• Analyze</li><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 06</li><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

# Collect

*Lab 03 – Collect: Connections  
Lab 05 – Collect: Virtualize*



# CPD Collect

## 1. Provision in-cluster databases

**Provision, host, and manage these data sources directly on the CPD cluster**

 CockroachDB <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Partner</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Premium</span> Ultra-resilient distributed SQL clusters designed for global business.	 Data Virtualization <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Enabled ✓</span> Query many data sources as one.	 IBMDb2 <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Premium</span> Relational database that delivers advanced data management and analytics capabilities for transactional and warehousing workloads.	 Db2 Event Store <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> In-memory data store capable of extremely high speed ingest and deep, real-time analytics.	 Db2 Warehouse <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Enabled ✓</span> Data warehouse designed for high-performance, in-database analytics.
 EDB Postgres <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Partner</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Premium</span> Object-relational database designed for developers.	 IBM Db2 for z/OS <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> Create databases in Db2 for z/OS and work directly with the data from IBM Cloud Pak for Data	 MongoDB Enterprise Advanced <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Partner</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Premium</span> Scalable, NoSQL database for enterprise deployments.	 Virtual Data Pipeline <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">IBM</span> <span style="border: 1px solid #ccc; border-radius: 50%; padding: 2px;">Premium</span> Access all the data you need for analytics and application testing without impacting production databases.	

# CPD Collect

## 2. Connect to existing data sources: IBM \*

\* Note: This list is constantly updated and shows what exists as of August 27, 2020.

**Connect directly to these data sources and perform the CPD component functionality shown**

IBM Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio
Analytics Engine HDFS				✓	✓
Classic Federation		✓			
Cloud object storage (IBM)		✓		✓	✓
Cloud object storage (infra)				✓	✓
Cloudant				✓	✓
Cognos Analytics				✓	✓
Compose for MySQL				✓	✓
Data Set		✓			
Data Virtualization	✓			✓	✓
Data Virtualization Mgr z/OS			✓		
PostgreSQL databases				✓	✓
Db2		✓	✓	✓	✓
Db2 Big SQL			✓	✓	✓
Db2 Event Store			✓	✓	
Db2 for i			✓	✓	✓
Db2 for z/OS		✓	✓	✓	✓
Db2 Hosted				✓	✓
Db2 on Cloud		✓	✓	✓	✓
Db2 Warehouse	✓	✓	✓	✓	✓
Db2 Warehouse on Cloud	✓	✓	✓	✓	✓
Distributed Transactions		✓			
DRS		✓			

IBM Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio		
External Source				✓			
External Target				✓			
HDFS via Hadoop					✓	✓	
Hierarchical				✓			
Hive via Hadoop					✓	✓	
Impala via Engine for Hadoop							
Informix				✓	✓	✓	✓
Informix Enterprise / Load				✓			
ISD Input / Output				✓			
Java Integration				✓			
Lookup File Set				✓			
Netezza				✓	✓		
Planning Analytics					✓	✓	
Obj. Strg. OpenStack Swift					✓	✓	
PureData for Analytics					✓	✓	
WebSphere MQ				✓			
Z/os DVM sources (VSAM, IMS, Adabas, etc.)				✓			

# CPD Collect

## 2. Connect to existing data sources: Third-party \*

\* Note: This list is constantly updated and shows what exists as of August 27, 2020.

**Connect directly to these data sources and perform the CPD component functionality shown**

Third-party Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio	Third-party Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio
Amazon Redshift			✓	✓	✓	Looker				✓	✓
Amazon S3		✓		✓	✓	MariaDB			✓		
Apache Cassandra	✓					MSFT Azure Blob and File		✓			
Apache Derby			✓			MSFT Azure Lake Store				✓	✓
Apache Hbase		✓				MSFT Azure SQL DB				✓	✓
Apache HDFS				✓	✓	MSFT SQL Server	✓	✓	✓	✓	✓
Apache Hive			✓	✓	✓	Minio				✓	✓
Apache Kafka	✓					Mongo			✓	✓	
Azure Storage	✓					MySQL			✓	✓	✓
BDFS	✓					ODBC		✓			
Cloudera Impala			✓	✓	✓	OData				✓	✓
Dropbox				✓	✓	Oracle		✓	✓	✓	✓
Filesystem	✓			✓	✓	Pivotal Greenplum		✓		✓	✓
FTP Enterprise	✓					PostgreSQL	✓		✓	✓	✓
FTP	✓			✓	✓	Salesforce.com		✓		✓	✓
Generic JDBC				✓	✓	SAP HANA			✓		
Google BigQuery	✓		✓	✓	✓	SAP Data Object		✓		✓	✓
Google Cloud Storage	✓			✓	✓	Snowflake		✓	✓	✓	✓
HDFS Generic web-HDFS				✓		Sybase Enterprise		✓	✓	✓	✓
HDFS HttpFS				✓		Sybase IQ / OC		✓		✓	✓
Hive JDBC	✓		✓			Tableau				✓	✓
Hive JDBC CDH	✓			✓		Teradata		✓	✓	✓	✓
Hive JDBC HDP	✓			✓							
Hortonworks HDFS				✓	✓						

# CPD Collect

## 2. Connect to existing data sources: Data Files

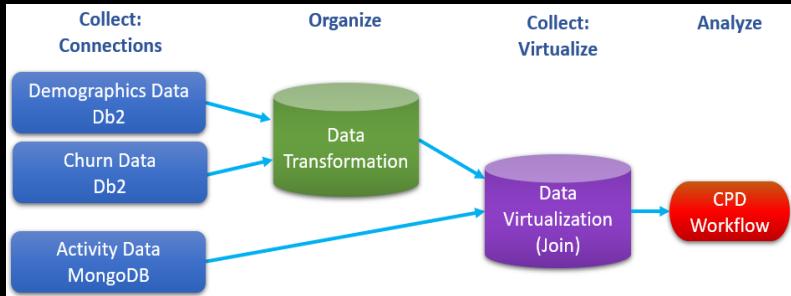
**Connect directly to these data sources and perform the CPD component functionality shown**

Other Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio
Comma Separated Value (CSV) files	✓	✓	✓		✓
Microsoft Excel Spreadsheets			✓		✓
Sequential File		✓			
Tab Separated Value (TSV) files			✓		

**Connecting to any of the data sources by service:**

Service	Comment
Cognos Dashboards	You can use the local and remote data sets that already exist in your analytics projects. You can create connections by selecting <i>Add data source</i> from the analytics dashboard menu.
DataStage Edition	You can transform data that is in a catalog by searching for the data that you want to use and selecting <i>Transform</i> .
Data Virtualization	You can create connections that can be used to virtualize data from the following locations: 1) <i>Connections</i> page, 2) <i>Data Sources</i> page in the Data Virtualization service
Watson Knowledge Center	You can create connections that can be used in the catalog and connections that can be used to curate data.
Watson Studio	Ideally, you should use data that is already in a catalog by searching for the data you want there and add it to an analytics project. Alternatively, you can create connections that can be used in analytics projects from the following locations: 1) <i>Connections</i> page, 2) <i>Assets</i> page of the analytics project. You can also <i>Add data</i> from files.

# Lab-3 Collect: Connect



## Review DB2 Data

- Customer Demographics
- Customer Churn
- Credentials

## Review DB2 Connection

- Connection Parameters

## Review MongoDB Data

- Customer Activity
- Credentials

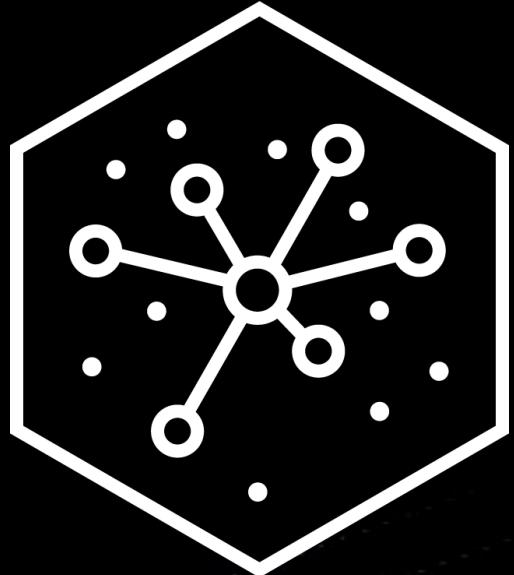
## Review MongoDB Connection

- Connection Parameters

# Cloud Pak for Data 3.0.1

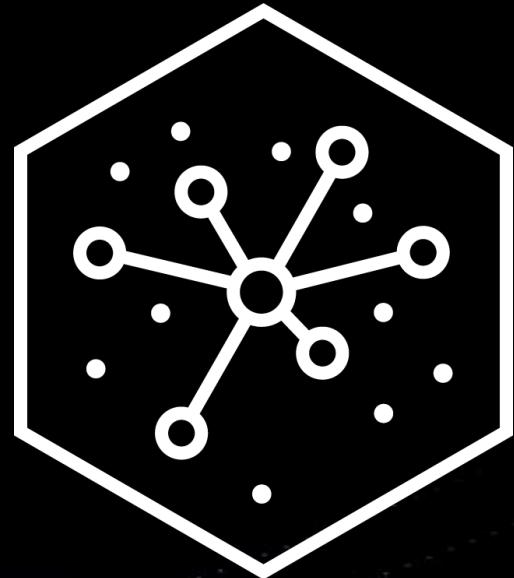
Work on Lab 3

Session restarts at xx AM



# Organize – Deeper Dive

*Lab 13 – Organize*





# CPD Organize

## Watson Knowledge Catalog (WKC)



**Data Scientists**  
**Data Analysts**  
**Business Analysts**

- Search and find relevant data
- Prepare data for consumption and analysis
- Consume and analyze the data
- Rate, comment on and share the data



**Data Stewards**  
**CDO**  
**LOB Risk**  
**LOB Product**

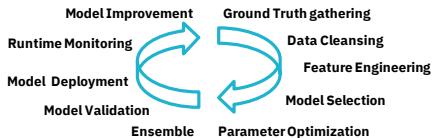
- Manage metadata repository
- Manage Reference Data
- Data governance workflow
- Discover metadata assets
- Classify data assets
- Data stewardship
- Build data glossary
- Data ownership
- Data lineage



**Quality Analysts**

- Profile data
- Understand, monitor and remediate data quality
- Apply validation rules

### AI Lifecycle



### Enterprise Data Consumption

### Enterprise Data Governance

### Enterprise Data Quality

**WKC is installed when you install any of the following services:**  
*Watson Studio, Watson Machine Learning, Watson OpenScale*

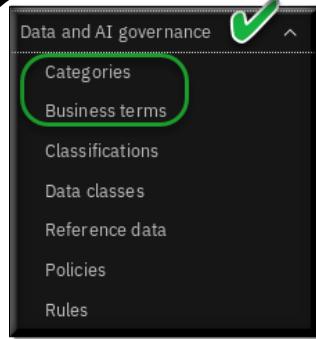
**IBM Watson Knowledge Catalog on Cloud Pak for Data**

End-to-End Platform for Business-Ready Data



# CPD Organize

## Data and AI governance: Categories and Business terms



**Categories** provide logical structure to a business glossary

**Business terms** standardize definitions of business concepts

1. Manually create Categories and Business Terms
2. Import Categories and Business Terms from CSV or XML files
3. Import a Glossary from an **industry accelerator**



The grid displays 16 different AI-powered business dashboards and analytics tools, each with a title, a brief description, and a timestamp of May 01, 2020, followed by the IBM logo. The tools include:

- Customer 360 Degree View
- Credit Card Fraud Analysis
- Contact Center Optimization
- Loan Default Analysis
- Healthcare Location Services Optimization
- Utilities Customer Attrition Prediction
- Streaming Analytics for Customer Life Event Prediction
- Customer Offer Affinity
- Customer Attrition Prediction
- Customer Segmentation
- Customer Life Event Prediction
- Utilities Demand Response Program Propensity
- Utilities Payment Risk Prediction
- Intelligent Maintenance Prediction

Each dashboard includes logos for Watson Knowledge Catalog, Cloud Pak for Data Industry, Watson Studio, and Watson Machine Learning.



# CPD Organize

## Data and AI governance: Policies and Rules

The sidebar shows a tree structure under 'Data and AI governance'. The 'Policies' node is highlighted with a green oval.

- Categories
- Business terms
- Classifications
- Data classes
- Reference data
- Policies (highlighted)
- Rules

**Policies** describe how to control data and consist of one or more rules

**Rules** describe the criteria for compliance with business objectives

The interface displays two main sections: 'Policies' and 'Rules', each with tabs for 'Published' and 'Draft'. A search bar and sorting options are also present.

**Policies**

- Data Privacy**  
Company-wide data privacy policy for se...  
[Data Privacy](#)  
Last modified: May 28, 2020
- Net Gains and Net Losses ar...**  
A customer can only have a value in Net...  
[Customer Churn Category](#)  
Last modified: May 28, 2020

**Rules**

- All Credit Card Information Must be Protected**  
All constructs of a credit card must be protected to ensure that those who should not be able to view the information are not allowed to. The information can be redacted since it typically is not used as a unique identifier. This includes the Credit...  
[Data Privacy](#)  
Governance rule | Last modified: May 28, 2020
- All Email Addresses Must be Protected**  
All Email Addresses must be protected to ensure that those who should not be able to view the information are not allowed to. The information must be masked (obfuscated) where the original format and validity of the email address is preserved...  
[Data Privacy](#)



# CPD Organize

## Data and AI governance: Classifications and Data classes

The sidebar menu includes:

- Categories
- Business terms
- Classifications** (highlighted with a green oval)
- Data classes** (highlighted with a green oval)
- Reference data
- Policies
- Rules

**Classifications** describes the sensitivity level of data

**Data classes** describe the contents of data in a column in a structured data set

**Classifications** interface:

- Published tab is selected.
- Search bar: Find classifications.
- Sort by: Name.
- Items listed:
  - Confidential**:  
Confidential data is data that if compromised in some form, is likely to result in significant and/or long-term harm to an individual whose data it is. Access to confidential information is restricted to those who have a legitimate business need to know.  
[uncategorized]  
Last modified: May 27, 2020
  - Personally Identifiable Information**:  
Personally identifiable information (PII) is defined as any data that could potentially identify a specific individual or organization, and which can be used to distinguish one person from another. This includes things like names, addresses, email addresses, telephone numbers, and social security numbers.  
[uncategorized]  
Last modified: May 27, 2020

**Data classes** interface:

- Published tab is selected.
- Search bar: Find data classes.
- Items listed:
  - Account Number**:  
A value representing an Account Number.  
[uncategorized]  
Last modified: May 27, 2020
  - Address Line 3**:  
Address Line 3 of a multi-line address.  
[uncategorized]  
Last modified: May 27, 2020



# CPD Organize

## Data and AI governance: Reference Data

The sidebar menu includes:

- Categories
- Business terms
- Classifications
- Data classes
- Reference data** (highlighted with a green oval)
- Policies
- Rules

**Reference Data Sets** define list of permissible values that are allowed for use within a data field.

May be referenced by Business Terms, Policies, Rules and Data Classes

The page has tabs for **Published** and **Draft**. A search bar says **Find reference data**. The main list shows:

- State and Province Codes** (selected, highlighted with a gray box and an arrow pointing to the right)
- Customer Churn Category**

The page title is **State and Province Codes** with a **Published** status indicator. It has tabs for **Overview** and **Related content**. A search bar asks **What are you looking for today?**. The table lists:

	<input type="checkbox"/> Code	↑	Value
▼	<input type="checkbox"/> AA		Armed Forces (the) Americas
▼	<input type="checkbox"/> AB		Alberta
▼	<input type="checkbox"/> AE		Armed Forces Europe
▼	<input type="checkbox"/> AK		Alaska



# CPD Organize

## Auto-discover assets

### Data Discovery

CUSTOMER_DEMOGRAPHICS			
DOB	100%	Date of Birth 100% ▾	—
ESTINCOME	100%	NoClassDetected 100% ▾	—
GENDER	100%	Gender 100% ▾	Gender 100% ✕
HOMEOWNER	100%	Indicator 100% ▾	Home Owner 100% ✕
ID	100%	NoClassDetected 100% ▾	—
LATITUDE	100%	Latitude 100% ▾	—
LONGITUDE	100%	Longitude 100% ▾	—
STATE	92%	US State Code 92% ▾	—
STATUS	100%	Code 100% ▾	—
TAXID	93%	US Social Security Num... 93% ▾	—
ZIP	92%	US Zip Code 92% ▾	—

Use machine learning based auto-discovery to:

- ① Analyze data quality
- ② Analyze columns (Classify data)
- ③ Assign Business terms

You can perform discovery with data sampling to allow for self-service data access with a search.



# CPD Organize

## Publish to a catalog

### Catalog and govern your assets

Catalogs / CPD Workshop Catalog

#### CPD Workshop Catalog

Browse Assets    Access Control    Settings

What assets are you looking for?

Any type    Any source    Any tag

Showing 8 of 8 items

<input type="checkbox"/> Name	Owner	Tags
<input type="checkbox"/> Customer Activity	CPD User	
<input type="checkbox"/> Customer Churn	CPD User	
<input type="checkbox"/> Customer Churn	CPD User	
<input type="checkbox"/> Customer Demographics	CPD User	
<input type="checkbox"/> Db2Warehouse	CPD User	global...

**Watson Recommend:** Highly Rated    Recently Added

Asset Type	Asset Name	Owner	Added	Reviews
Data asset	Warehouse	CPD User	May 28, 2020 10:38 AM	0 reviews
Data asset	Customer Demographics	CPD User	May 28, 2020 10:41 AM	1 review
Data asset	Customer Activity	CPD User	May 28, 2020 10:42 AM	0 reviews

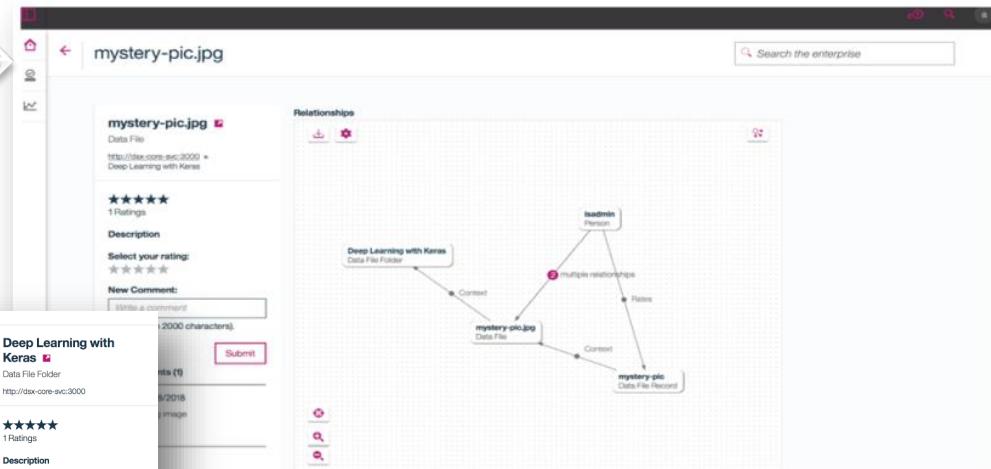
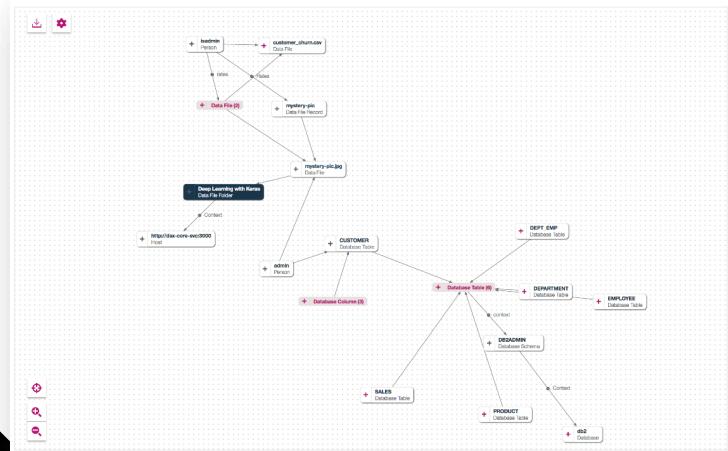
Showing 8 of 8 items



# CPD Organize

## Relationship graph with explorer

- Explore relationships between data assets, terms, analytic assets, users, etc.
- Gain in-depth understanding of metadata through crowdsourcing (e.g., ratings, comments) and machine learning



Explore deeper to understand context and usage patterns



# CPD Organize

## Refine data with visualizations

**Refine** can cleanse and shape tabular data with a graphical flow editor using functions and logical operators.

Use it to remove data that is incorrect, incomplete, improperly formatted, etc.

Shape the data by filtering, sorting, combining or removing columns. You can create a Data Refinery flow as a set of ordered operations on the data to run repeatedly any time.



ID Smallint	GENDER String	STATUS String	CHILDREN Smallint	ESTINCOME Decimal	HOMEOWNER String	AGE Smallint	TAXID String
Identif...	Gender	Code	Code	Not clas...	Indicator	Code	US So...
481	F	M	2	28267	N	30	386283240
482	F	M	2	36725.1	N	56	162447113
483	M	S	1	94188.3	N	58	673845765
484	F	M	2	91861	Y	42	209619292



Data Refinery also includes a graphical interface to profile data to validate it with 20+ customizable charts that give perspective and insights into the data.



# CPD Organize

## Profile data

The **Profile** of a data asset includes generated metadata and statistics about the textual content of the data.

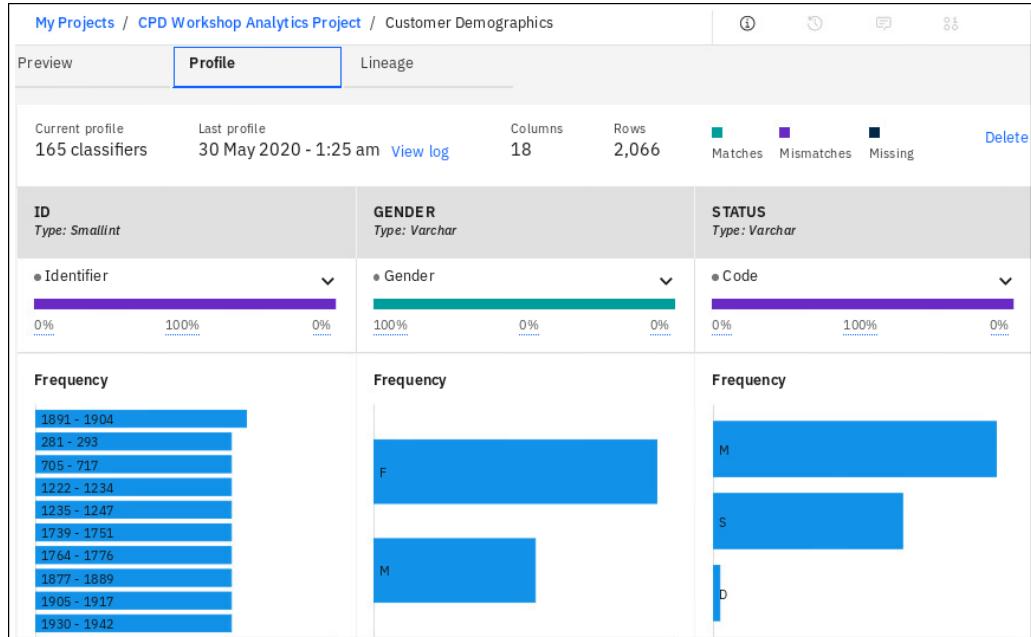
All catalog or project members can see data asset profiles.

Profiles are automatically created:

- In catalogs, profiles for unstructured data assets are created automatically, regardless of whether policies are enforced
- In governed catalogs, profiles for structured data assets are created automatically

Profiles can be manually created:

- In ungoverned catalogs for structured data assets
- In projects for both structured and unstructured data assets



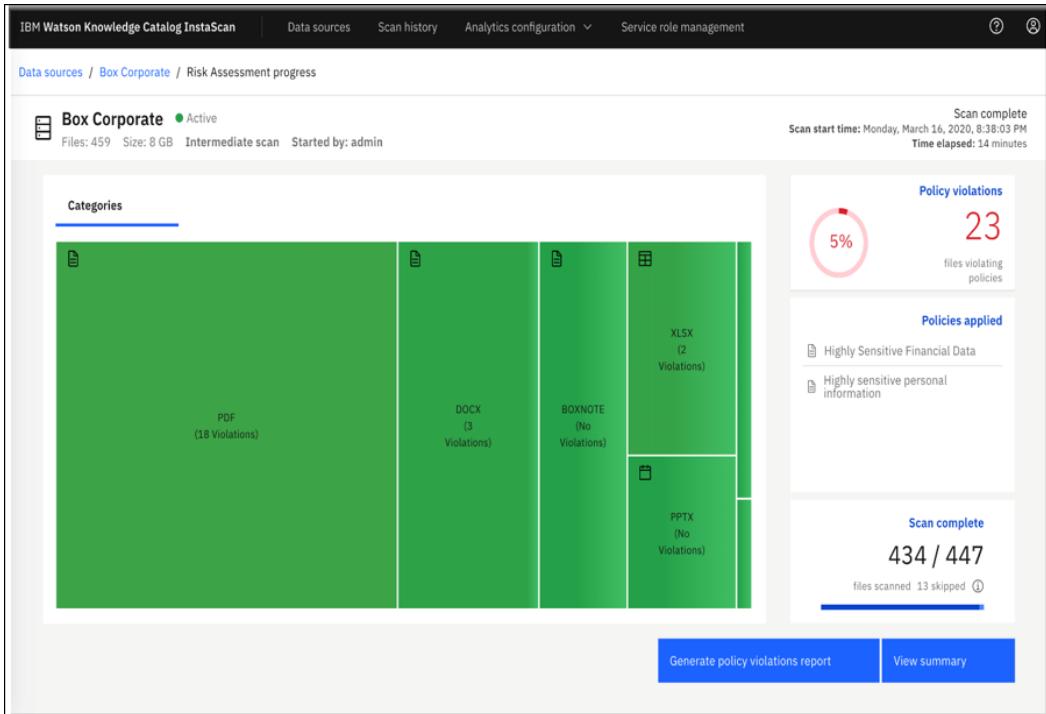


# CPD Organize

## InstaScan

**InstaScan** can perform risk assessments and compliance checks of unstructured data

- Scan email, PDFs, word processor documents and images
- Quickly determine which areas have high concentration of sensitive information and prioritize hot spots
- Automatically apply classification labels to data that violates corporate policies
- Create and share ongoing compliance reports with CISO or regulators
- Integrates with Box Shield to provides a comprehensive data privacy solution for unstructured data in Box





# CPD Organize

## Search for data

The screenshot shows the CPD Organize interface with a search bar at the top containing the query "churn". Below the search bar is a "Suggestions" section with a "churn" link. The main area displays a title "Search results for churn" and filtering options: "Any type", "Any tag", and "Steward/Owner". It shows "Showing 19 of 19 items" and lists three assets:

Name	Type
Gender Categories > Customer Churn Category Customer Churn Category	Business term
Customer Churn All catalogs > CPD Workshop Catalog	Data asset
Customer Churn All projects > CPD Workshop Analytics Project	Data asset

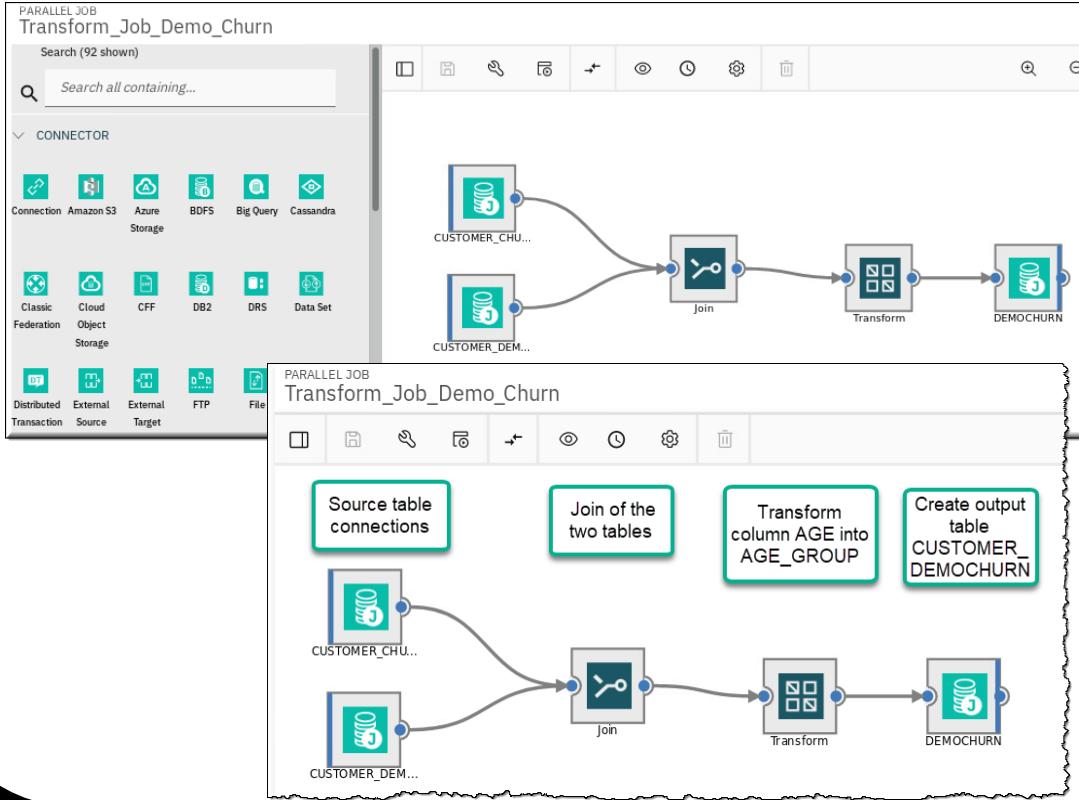
### Relevancy of search results factors:

<i>Text match</i>	The provided text is searched in the asset name & description, where name contributes more to the higher place in the result list.
<i>Asset rating</i>	The higher the average rating the asset has, the higher it is on the results list.
<i>Comments</i>	The higher the number of comments, the higher the asset is on the results list.
<i>Context match</i>	The search results list might contain the closest neighbors of assets that are returned based on the text match.
<i>Modification date</i>	The assets that were modified recently are more likely to be returned in the search results.
<i>Quality score</i>	The higher the score, the higher the asset is on the results list. Quality score applies to database tables, views & columns, design tables, views & columns, data file records & fields.
<i>Usage</i>	The more relationships of type uses an asset has, the higher it is on the results list.



# CPD Organize

DataStage - Transform and migrate data, build and execute ETL jobs at scale



- Use powerful data transformation capabilities
- ML infused Smart Job Clustering, Smart Job Assist and automated job sequencing
- Design once, run on any-cloud using Kafka
- Remotely execute job for co-located access to data, satisfy geopolitical requirements and save costs
- In-line data quality and governance to build trusted data when the data is being delivered to a target environment such as a data lake
- Create CI/CD pipelines with GitHub

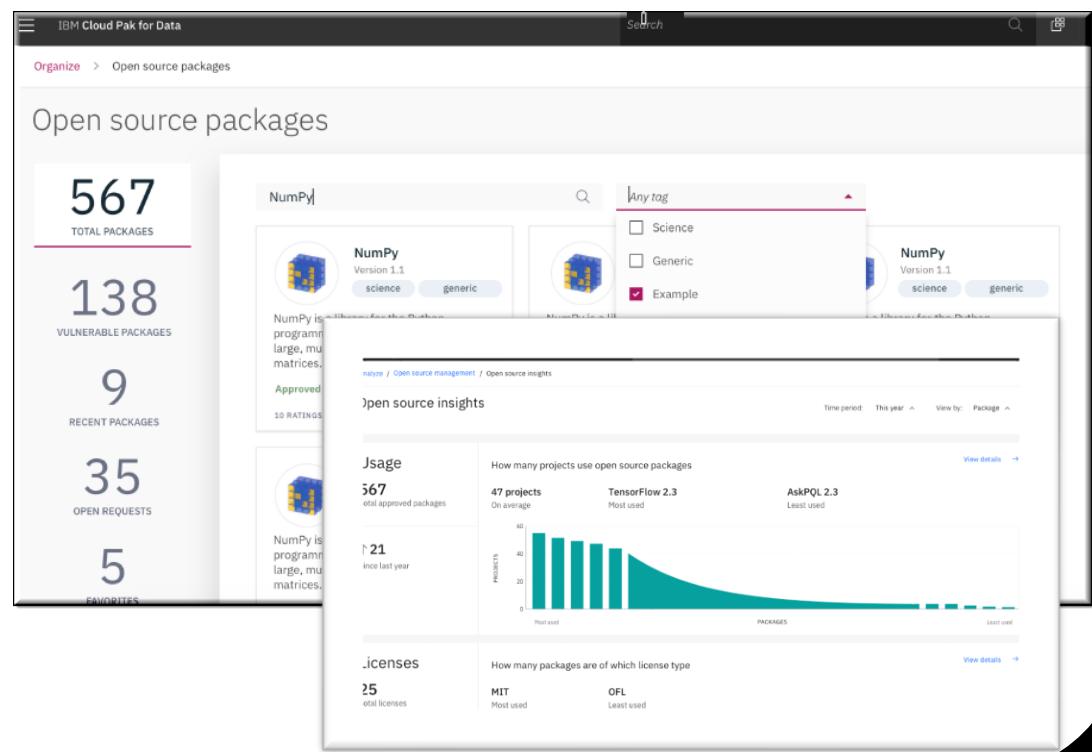


# CPD Organize

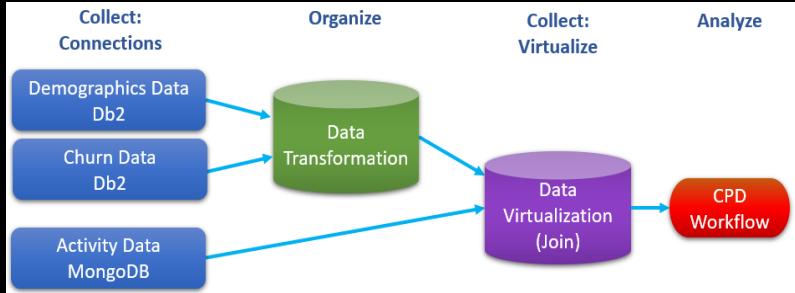
## Open Source Software Management

### Open Source Software Management is:

- A *centralized inventory* of approved open source packages to ensure code quality & security
- A vehicle for *self-service open source consumption* and workflow requests for developers
- A *correlation of vulnerabilities* across open source portfolio: highlight security risks
- An *infusion of collaboration* by allowing developers to rate and comment on opensource packages
- A *set of dashboards* for CIOs to manage risks and accelerate innovation with OSS



# Lab 13 – Organize- Deeper Dive



## Create a Project

- ✓ Add a Global DB2 Connection
- ✓ Add Connected Data
  - ✓ Customer Demographics Table
  - ✓ Customer Activity Table
- ✓ Preview Customer Demographics
- ✓ Profile Customer Demographics
  - ✓ Data Column Frequencies
  - ✓ Data Column Classification (e.g. TaxID-> US SSN)
- ✓ Refine Customer Demographics
  - ✓ Data Visualization
  - ✓ Data Shaping (Zip Code Padding)
  - ✓ Run Data Refinery Job

## Review Governance Artifacts

- ✓ Categories and Business Terms
- ✓ Classifications
- ✓ Data Classes
- ✓ Reference Data

## Data Protection Rules and Masking

- ✓ Create Policy
- ✓ Define Protection Rule
  - ✓ Data Class DOB masked for user=='developer'
- ✓ Assign the rule to the policy
- ✓ Demonstrate masked DOB on display for 'developer'

## Data Discovery Automation

- ✓ Crawl 2 DB2 Tables
- ✓ Classify Data Columns
- ✓ Score Data Columns, Data Table
- ✓ Map Business Terms to Data Columns
- ✓ Analysis Results in Workspace

## Data Transformation

- ✓ Review a Transformation Job
- ✓ Create a Transformation Job
- ✓ Use the Visual Palette
- ✓ Compile, Run Job, View Output

## Data Lineage

- ✓ Information Assets
- ✓ View Customer Demographics Data Lineage

# Lab 13 – Organize – Deeper Dive – Review

Preview Data – Ensure that you are looking at the correct (and useful) data

The screenshot shows the IBM Cloud Pak for Data interface. The top navigation bar includes 'IBM Cloud Pak for Data', 'My Projects / Data Analysis Project / CUSTOMER\_DEMOGRAPHICS', 'All' dropdown, and a search bar. Below the navigation is a tabs section with 'Preview' (selected), 'Profile', and 'Lineage'. A status bar indicates 'Schema: 18 Columns' and 'Preview: 1000 rows'. The main area displays a table of 18 columns and 18 rows of data. The columns are labeled: ID, GENDER, STATUS, CHILDREN, ESTINCO..., HOMEOW..., AGE, TAXID, CREDITC..., DOB, ADDRESS..., ADDRESS..., CITY, STA. The data rows show various demographic details for different individuals, such as gender (M/F), status (S/M), number of children (0-2), estimated income (e.g., 38000, 5237.63), home ownership (N/Y), age (24-61), tax ID, credit card number, date of birth, address, city, state, and zip code.

ID Smallint	GENDER String	STATUS String	CHILDREN Smallint	ESTINCO... Decimal	HOMEOW... String	AGE Smallint	TAXID String	CREDITC... Char	DOB Date	ADDRESS... String	ADDRESS... String	CITY String	STA Char
0	F	S	1	38000	N	24	147889187	6549061697939	1947-11-11	159 HUTTON ST I		ABSECON	NJ
1	M	M	2	29616	N	49	113772166	6436360484417	1992-03-17	31 WOODLAND R		SAINT LOUIS	MO
2	M	M	0	19732.8	N	51	132420919	4849378808118	1907-09-08	1910 COCHRAN I		KEARNY	NJ
3	M	S	2	96.33	N	56	700548452	2926742654852	1980-04-29	187 HAYES MILL		RUSTON	LA
4	F	M	2	52004.8	N	25	141013706	4132500804622	1979-01-16	RR 1 BOX 57B		MONTGOMERY	AL
5	M	M	2	53010.8	N	19	163371244	2231773884473	1992-12-06	7850 45TH AVE N		CHESTER	MA
6	M	M	1	75004.5	N	65	182544864	7349439804241	1911-04-24	RR 1 BOX 47		NEW CASTLE	PA
7	M	M	0	19749.3	N	60	206227068	5553618912566	1912-11-23	515 KENSINGTO		ISSAQAH	WA
8	M	S	1	57626.9	Y	44	131059071	9119007527242	1916-04-26	6077 STATE ROU		SHAVERTOWN	PA
9	M	M	2	20078	N	33	119762649	6813572896826	1977-02-09	188 W OLYMPIC I		EL PASO	TX
10	F	M	2	47902	N	26	817366094	2046608099384	1905-02-01	21579 LARAMIE		PHILADELPHIA	PA
11	M	M	1	7545.96	Y	17	451541224	4045479553572	1924-10-01	1 PLAINVILLE CI		HAVERHILL	MA
12	F	S	0	78851.3	N	48	124158559	8979358234254	1933-06-14	4716 SW VIOLA C		TENAFLY	NJ
13	F	S	1	17540.7	Y	63	163930462	9050922700714	1942-12-12	1 HAUN RD		PHAROAH	OK
14	F	M	0	83891.9	Y	61	165912006	6325263828540	1947-01-25	4220 BARDSTOW		ARENA	ND
15	F	M	2	28220.8	N	39	235315405	4562971044212	1946-02-20	10167 OAK HOLL		STRAUSSTOWN	PA
16	F	S	0	28589.1	N	16	730825728	8111200911782	1916-10-05	D18 CALLE 5		HOUSTON	TX
17	F	M	2	5237.63	N	49	908144755	1349244456282	1938-10-02	3579 N 47TH AVI		KENOSHA	WI

# Lab 13 – Organize – Deeper Dive – Review

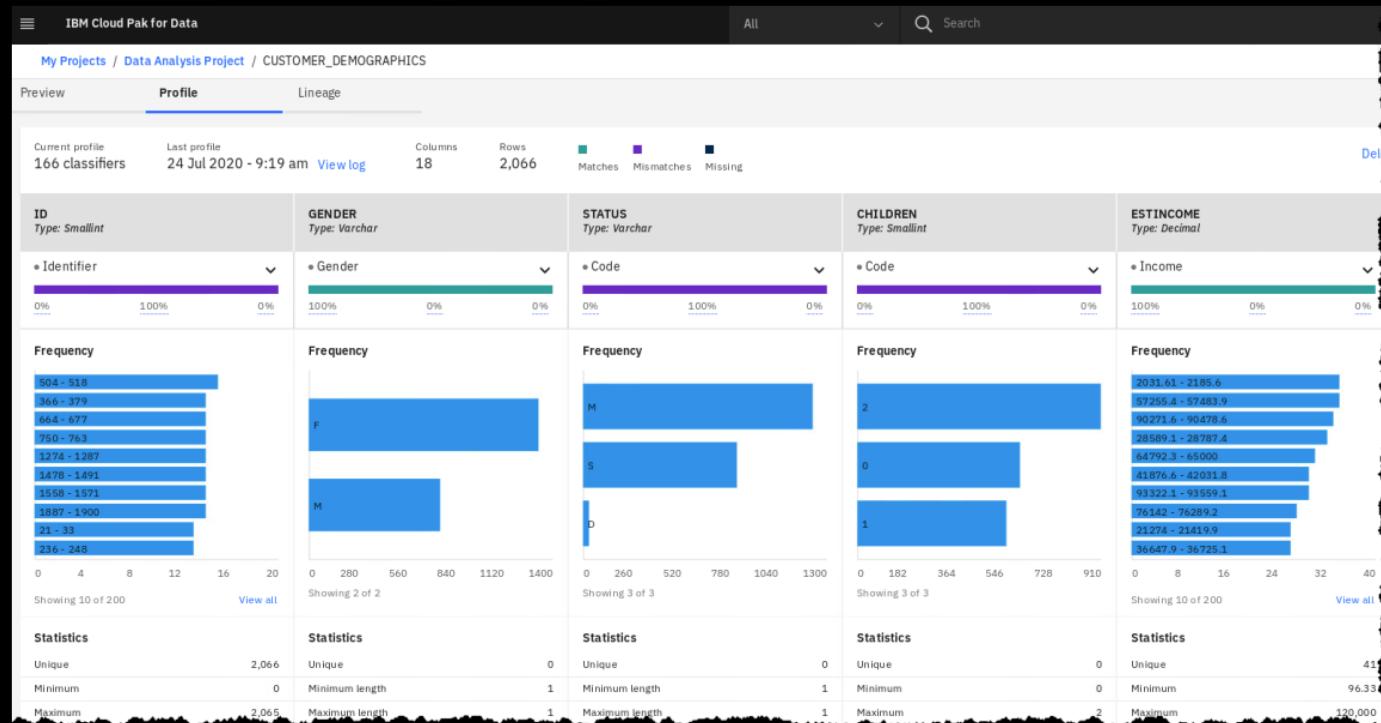
Data Classes – Classifiers for data. Many “out of the box”, but you can also create your own

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below it, the path 'My Projects / Data Analysis Project / CUSTOMER\_DEMOGRAPHICS' is visible. The main area has tabs for 'Preview', 'Profile' (which is selected and highlighted with a blue border), and 'Lineage'. A large icon of a document with a bar chart is centered. Below the icon, the text 'Data profile not yet created' is displayed, followed by the instruction 'You can create a profile from the first 5000 rows of the data set.' Two buttons are present: 'Select data classes' with a checkmark icon and 'Create profile'.

A modal dialog box titled 'Data classes' is shown. It includes a 'Clear all' button, a search bar, and a status message '166 of 166 selected'. The main area is labeled 'Available classifiers' and lists several options, each with a checked checkbox. The listed classifiers are: DUNS Number, Ethnicity, Honorific, Country Name, US Phone Number, and Political Party. At the bottom of the dialog are 'Cancel' and 'Apply' buttons, with 'Apply' being highlighted in blue and featuring a green checkmark icon.

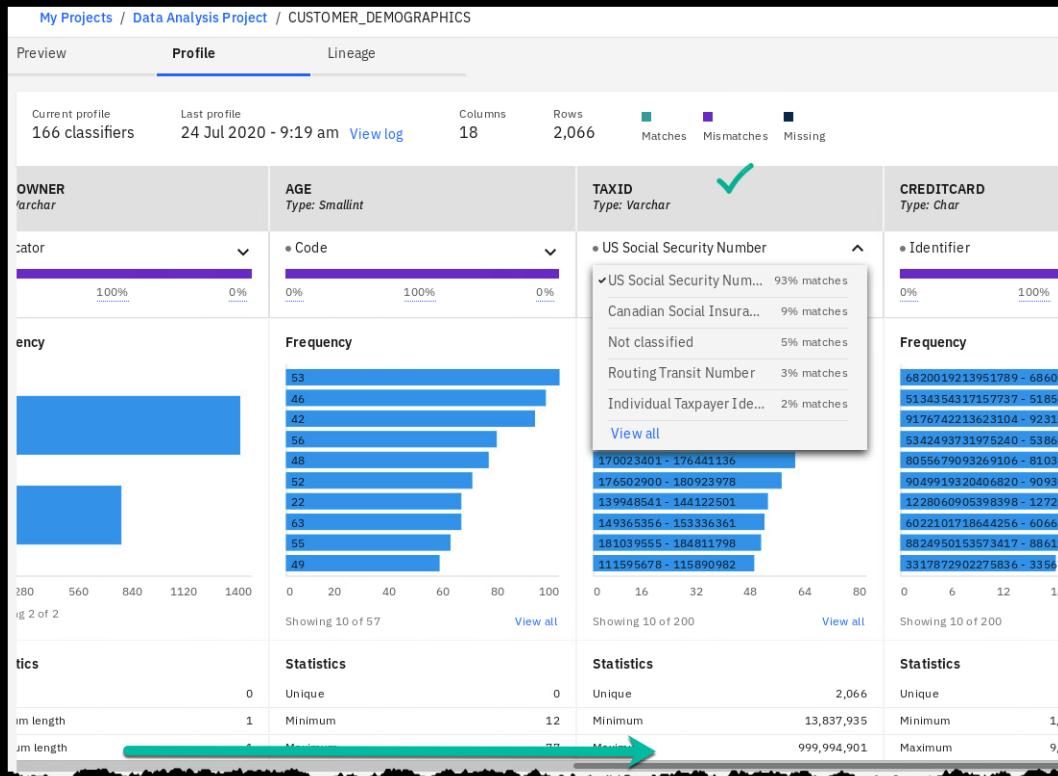
# Lab 13 – Organize – Deeper Dive – Review

Data Profiling – Quick view of data anomalies, quality, frequency...



# Lab 13 – Organize – Deeper Dive – Review

Data Profiling – ...and ‘automatic’ linkage to different data classes (even in same field)



# Lab 13 – Organize – Deeper Dive – Review

Data Shaping – can be used for ‘self service’ data shaping operations...

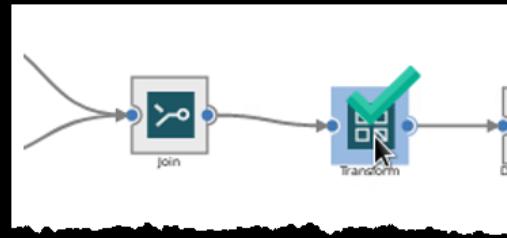
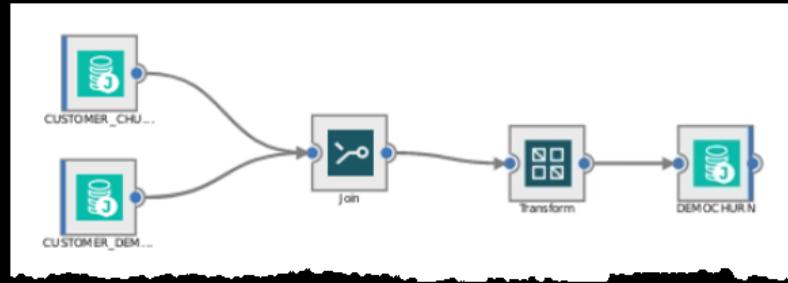
The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below it, the project path is 'My Projects / Data Analysis Project / CUSTOMER\_DEMOGRAPHICS / Refine data'. A blue button labeled 'Operation +' is visible. The main area has tabs for 'Data', 'Profile', and 'Visualizations', with 'Data' selected. A text input field says 'Code an operation to cleanse and shape your data'. Below is a table with columns: ID (Integer), GENDER (String), STATUS (String), CHILDREN (Integer), ESTINCOME (Decimal), HOMEOWNER (String), and AGE (Integer). The table contains 12 rows of customer demographic data.

ID	GENDER	STATUS	CHILDREN	ESTINCOME	HOMEOWNER	AGE
1	F	S	1	38000	N	24
2	M	M	2	29616	N	49
3	M	M	0	19732.8	N	51
4	M	S	2	96.33	N	56
5	F	M	2	52004.8	N	25
6	M	M	2	53010.8	N	19
7	M	M	1	75004.5	N	65
8	M	M	0	19749.3	N	60
9	M	S	1	57626.9	Y	44
10	M	M	2	20078	N	33
11	F	M	2	47902	N	26
12	M	M	1	7545.96	Y	17

# Lab 13 – Organize – Deeper Dive – Review

Note that you are switching between capabilities seamlessly because they are on the same platform...

No need to spend time/effort “integrating the integration tools together”, more time working with the data and result sets



# Lab 13 – Organize – Deeper Dive – Review

There can be several methods to transform the data for a use case.

The Transformer stage contains a derivation builder - one way to create logic to your use case.

	LINK_52.ZIP4	ZIP4
	Link_52.LONGITUDE	LONGITUDE
	Link_52.LATITUDE	LATITUDE
	If Link_52.AGE < 18 THEN "Child" ELSE IF Link_52.AGE < 30 THEN "Young adult" ELSE IF Link_52.AGE < 65 THEN "Adult" ELSE "Senior"	AGE_GROUP
		✓

Runtime column propagation

## Derivation Builder - AGE\_GROUP VARCHAR(11)

### Derivation

```
If Link_52.AGE < 18 THEN "Child" ELSE IF Link_52.AGE < 30 THEN "Young adult" ELSE IF Link_52.AGE < 65 THEN "Adult"  
ELSE "Senior"
```



# Lab 13 – Organize – Deeper Dive – Review

## Automated Discovery – Assess quality of data

The screenshot shows the 'Automated discovery job' configuration interface. At the top, it says 'Connection \*' with 'Db2 Advanced Edition' selected, indicated by a green checkmark. Below that is the 'Discovery root' field, which contains the example value 'Example: schema[db\_name|schema\_name];table[db\_name|schema\_name|table\_name]'. A 'Browse' button is highlighted with a pink box and a green checkmark. Under 'Discovery options', there are four checkboxes: 'Analyze columns', 'Analyze data quality', 'Assign terms', and 'Publish results to catalog'. The first three are checked, while 'Publish results to catalog' is checked and highlighted with a pink box and a green checkmark.

The screenshot shows the 'Automated discovery' configuration summary. It includes a large green checkmark icon. Below it, the title 'Automated discovery' is followed by a detailed description: 'Choose automated discovery if you want to see the details about the quality of your data based on an in-depth analysis of all assets. The source assets are imported. Automated discovery is suitable for smaller data sources, or selected components of larger data sources.' To the right, a summary of the selected 'Discovery options' is shown:

Discovery option	Status
Analyze columns	Selected (checked)
Analyze data quality	Selected (checked)
Assign terms	Selected (checked)
Publish results to catalog	Selected (checked)
Use data sampling	Not selected (unchecked)

# Lab 13 – Organize – Deeper Dive – Review

## Automated Discovery – Assess quality of data

The screenshot displays two main sections of the IBM Cloud Pak for Data interface.

**Top Section (Organize-Workspace):**

- Dashboard:** Shows a circular progress bar for "Data quality threshold" at 2 (100%), a histogram for "Quality score distribution" peaking around 80, and four bar charts for "Analysis", "Relationships", "Primary keys", and "Relationships".
- Project Details:** Shows 2 data assets, 0 PII data assets, 0 reviewed data assets, 1 connection, 0 critical data issues, created by "cpouser" on Jun 30, 2020, and last modified on Jun 30, 2020.
- Rule Run Status:** Shows "No data available" with a status icon.
- Top 5 Selected Classes:** A bar chart showing counts for Customer, Employee, Product, Region, and US\_Social\_Security\_Number.

**Bottom Section (Data assets):**

- Data Asset Details:** Two cards for "CUSTOMER\_ACTIVITY" and "CUSTOMER\_DEMOGRAPHICS". Both have 99% and 97% completion rates respectively, with green checkmarks. They show connection details (jdbc:db2://worker5.clusterw9:32030/BLUDB: BLUD B.SOLUTIONS), import and publish dates (Jun 30, 2020, 13:00), and data analysis thresholds (80% for both).

# Lab 13 – Organize – Deeper Dive – Review

Data Protection rules  
are tied to Policy

The screenshot shows the 'Create a new rule' interface. At the top, it says 'IBM Cloud Pak for Data'. Below that, 'Create a new rule' is displayed. A message asks, 'Which type of rule do you want to create? Select one.' Two options are shown: 'Data protection rule' and 'Governance rule'. The 'Data protection rule' option is selected, indicated by a green checkmark next to its icon (a fingerprint). The 'Governance rule' option is shown with a blue icon (a building).

**Data protection rule**  
A rule to mask or deny access to data.

**Governance rule**  
A rule used as documentation.

The screenshot shows the 'Create new Policy' dialog box. It has fields for 'Policy name' (containing 'Mask PII Data' with a green checkmark), 'Primary category' (set to '[uncategorized]' with a 'Change' button), and 'Description (optional)' (with a text area labeled 'Type description here'). At the bottom, there are 'Cancel' and 'Save as draft' buttons, with 'Save as draft' being highlighted in blue and having a green checkmark.

Create new Policy

Policy name

Mask PII Data ✓

Primary category ⓘ

[uncategorized] Change

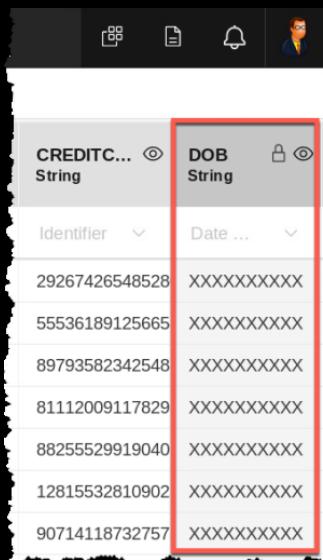
Description (optional)

Type description here

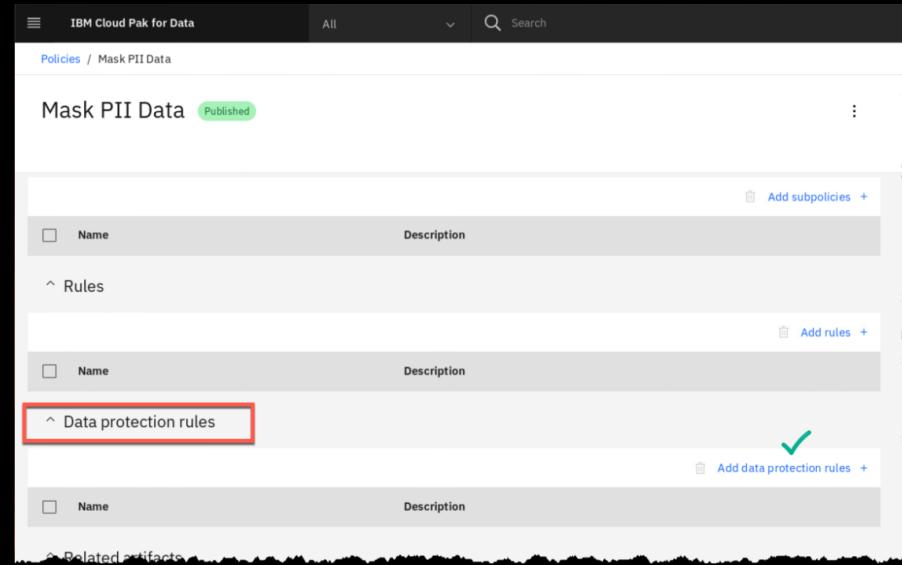
Cancel Save as draft ✓

# Lab 13 – Organize – Deeper Dive – Review

Note: This is ‘masking on the glass’.



CREDITC...	DOB
Identifier	Date ...
29267426548528	XXXXXX
55536189125665	XXXXXX
89793582342548	XXXXXX
81112009117829	XXXXXX
88255529919040	XXXXXX
12815532810902	XXXXXX
90714118732757	XXXXXX



Policies / Mask PII Data Published

Mask PII Data

Add subpolicies +

Name	Description
------	-------------

^ Rules

Add rules +

Name	Description
------	-------------

^ Data protection rules

Add data protection rules +

Name	Description
------	-------------

^ Related artifacts

# Lab 13 – Items of note (outside the labs)

Metadata Asset Manager is used to import certain types of metadata assets into the metadata repository (such as logical or physical data models, Business Intelligence models, etc).

Provides a managed process for comparing assets and ensuring you are not re-ingesting structures

The screenshot shows a web browser window displaying the IBM Knowledge Center. The URL is https://www.ibm.com/support/knowledgecenter/SSZJPZ\_11.7.0/com.ibm.swg.im.iis.mml.doc/101. The page title is "Overview of InfoSphere Metadata Asset Manager". The left sidebar contains a table of contents for "IBM InfoSphere Information Server Version 11.7.1 documentation", including sections like "Accessing the product documentation", "Reading command-line syntax", "Product accessibility", "Contacting IBM", "Notices and trademarks", and "Managing metadata". The main content area discusses the use of the Metadata Asset Manager to import, export, and manage metadata assets. It mentions that users can analyze assets, use them in jobs, assign them to terms, or designate stewards for the assets. The right sidebar features a "Bookmarks" section with various links to IBM products and services.

# Lab 13 – Items of note (outside the labs)

You can ‘rate’ and ‘comment’ on assets

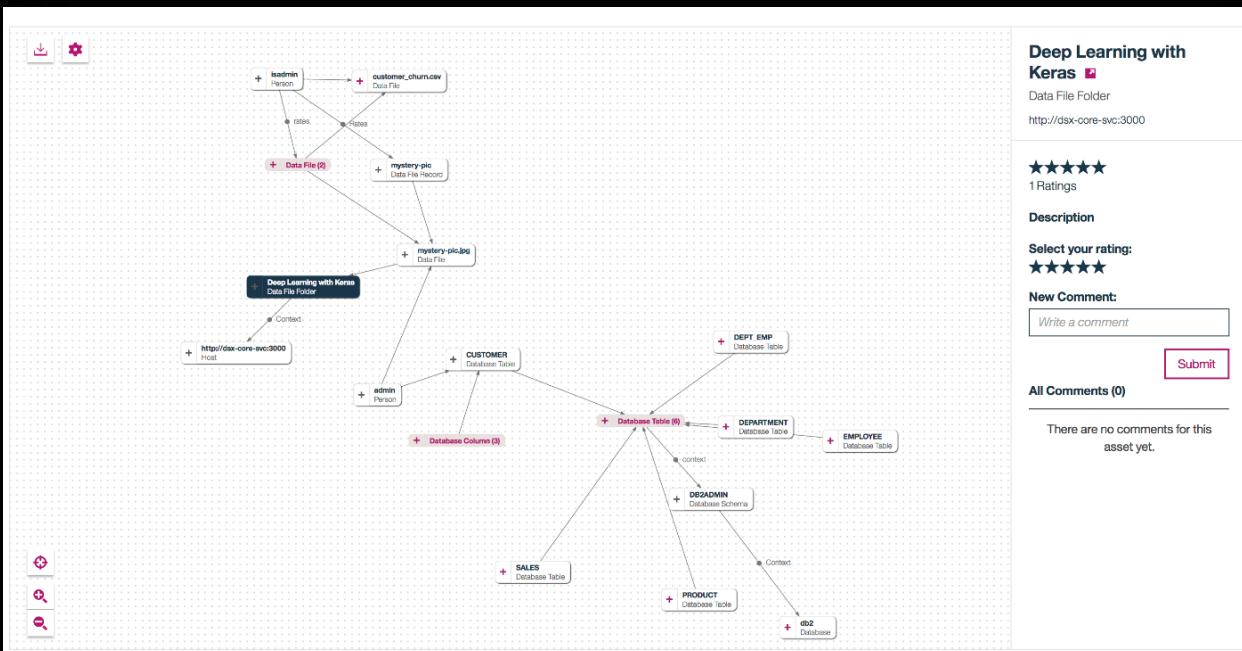
The screenshot shows the CPD Workshop Catalog interface. At the top, there's a navigation bar with 'Catalogs / CPD Workshop Catalog'. Below it, a search bar asks 'What assets are you looking for?' and filters for 'Any type', 'Any source', and 'Any tag'. A message 'Showing 8 of 8 items' is displayed. A table lists assets: Customer Activity, Customer Churn, Customer Demographics, and Db2Warehouse, each with a checkbox, owner (CPD User), and tags (global...).

A modal window titled 'Watson Recommend:' is open, showing three recommended assets:

- Warehouse**: Data asset, Owner: CPD User, Added: May 28, 2020 10:38 AM. It has 0 reviews.
- Customer Demographics**: Data asset, Owner: CPD User, Added: May 28, 2020 10:41 AM. It has 1 review (★★★★★).
- Customer Activity**: Data asset, Owner: CPD User, Added: May 28, 2020 10:42 AM. It has 0 reviews.

# Lab 13 – Items of note (outside the labs)

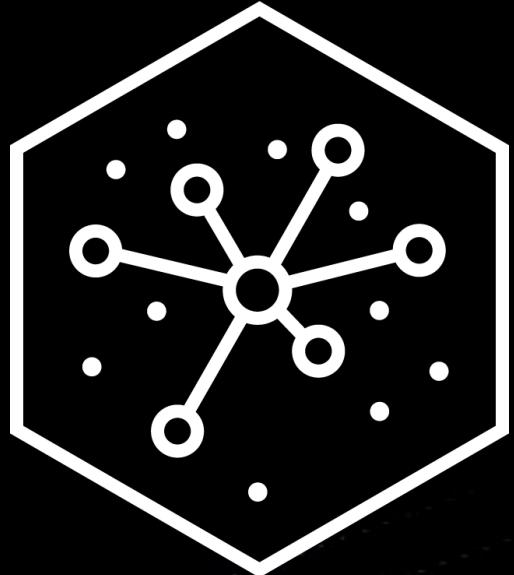
Relationship/Graph viewer allows you to graphically explore relationships between artifacts in your catalog



# Cloud Pak for Data 3.0.1

Work on Lab 13

Session restarts at 2 PM



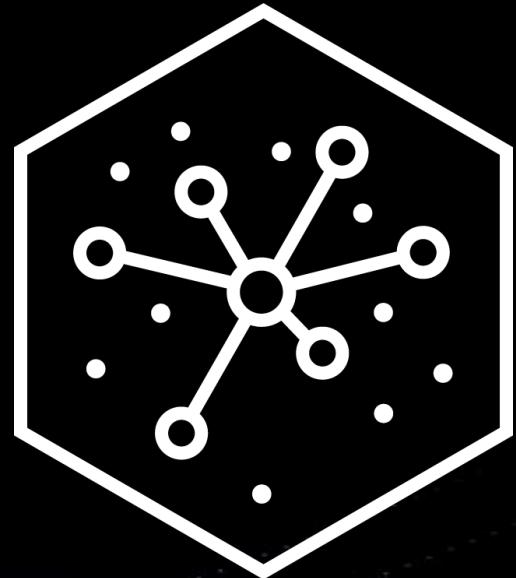
# IBM Analytics Modernization Workshop

## Part 2

	<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
	<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deeper Dive</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
	<ul style="list-style-type: none"><li>• Analyze</li><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 06</li><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

# Collect

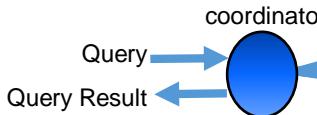
*Lab 05 – Collect: Virtualize*



# CPD Data Virtualization

## Constellation “Computational Mesh” benefit

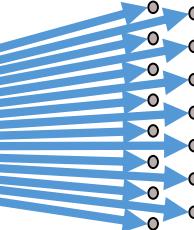
### Classic Federation & Edge Computing



Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes individually perform as much work as they can based on their own data. Individual results are sent back to the coordinator for final merging and remaining analytics.

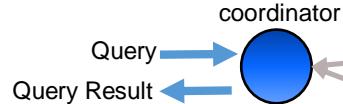


Coordinator receives intermediary results from all edge nodes, merges results, and performs remaining analytics

To be clear: Federation is a form of Data Virtualization and has been used successfully for many years in IBM products like Db2

CPD Data Virtualization uses a new Computational Mesh \* approach which meets the performance demands of today's modern data access requirements

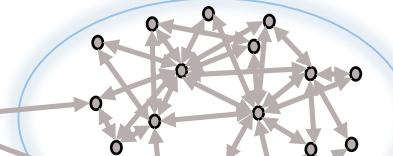
### New Computational Mesh



Query issued against the system

A coordinator receives the request and fans the work out to edge nodes

Edge nodes self organize into a constellation where they can communicate with a small number of peers. Nodes collaborate to perform almost all analytics, not only analytics on their own data.

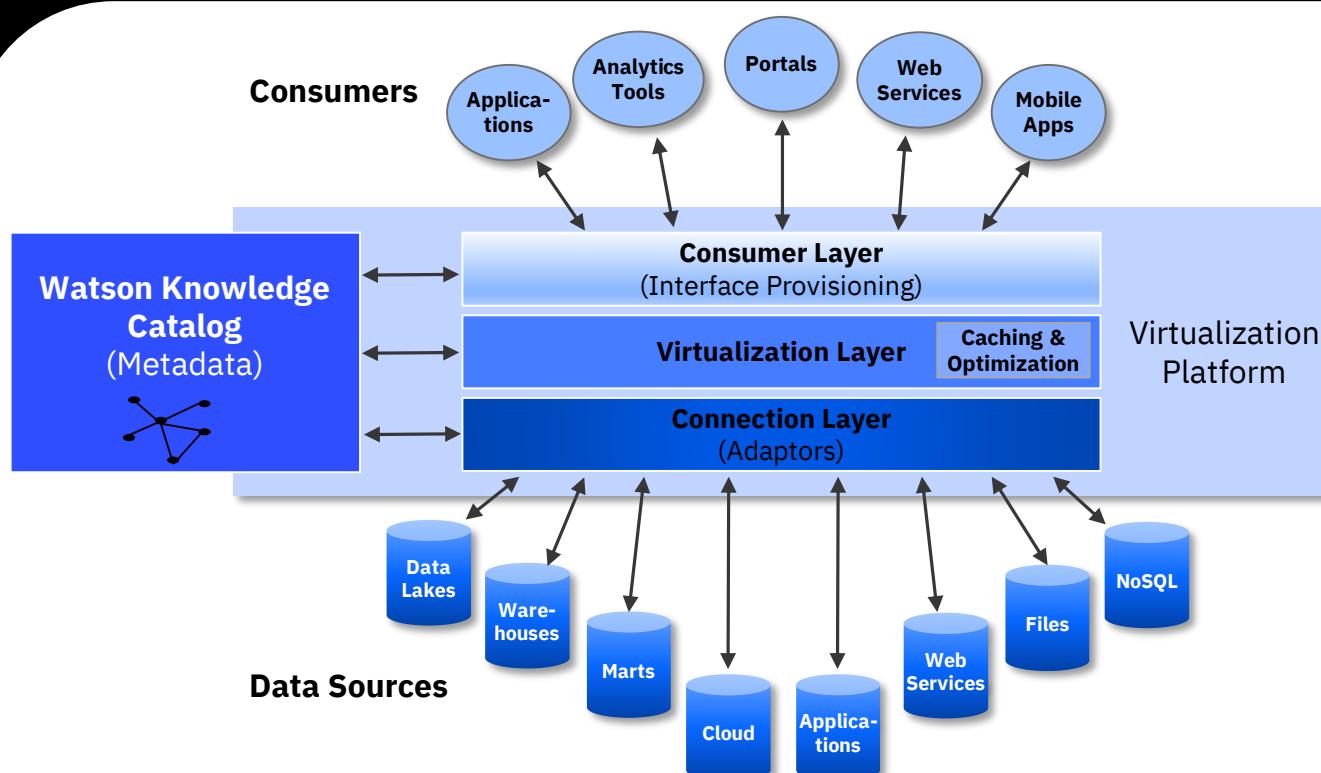


Coordinator receives mostly finalized results from just a fraction of nodes. Completes the final work for the query result.

\* Note: this is a work in progress. Remote Connectors with data source support is available today.

# CPD Data Virtualization

With Watson Knowledge Catalog (WKC) built in



- Provides the ability to search, view, access, manipulate, and analyze data
- No need to know or understand its physical format or location
- No need to move or copy it

# CPD Data Virtualization

## Benefits and use cases

### Benefits

#### **Simple:**

- Self-discovering, self-organizing cluster
- Joins provide a one source input to analytics

#### **Flexible:**

- Once established, it is easy to add new sources to the constellation
- Integrates disparate data assets with simple automation, providing seamless access to data as one

#### **Scalable:**

- Can access thousands of sources, IOT and edge devices

#### **Cost Effective:**

- Leverages the compute resources of source systems to execute the SQL

#### **Secure:**

- Inherits privileges & masking policies of the data sources
- Built in governance, security, and access control

### Use Cases

#### **Data Scientists:**

- Significant productivity increase getting access to sources discovery and assembly of data sets

#### **Current State answer requirement:**

- Current state required for up-to-date analytics
- One time access to data, then throw it away
  - e.g. “How much cash is ‘on hand’ across our branches worldwide?” “What is our current ‘claims’ liability?”

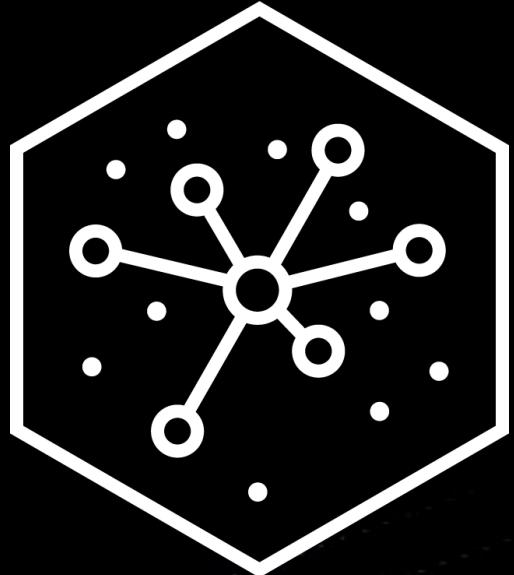
#### **ETL and/or Data Governance saturation**

- Self-service – In the event that Data Engineers cannot keep up with business demands for access to data

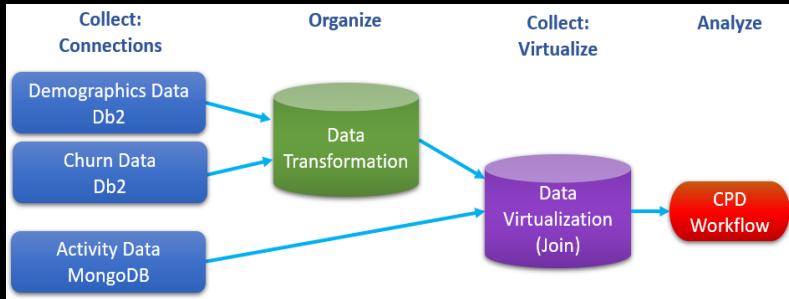
# Cloud Pak for Data 3.0.1

Work on Lab 5

Session restarts at xx AM



# Lab-5 Collect: Virtualize Review



Add Data Sources to Virtualization Facility

- ✓ DB2, MongoDB

Virtualize CUSTOMER\_DEMOCHURN

- ✓ Select Table
- ✓ Add to Cart
- ✓ Virtualize in Project

Virtualize ACTIVITY01

- ✓ Select Table
- ✓ Add to Cart
- ✓ Virtualize in Project

Join Virtual Tables

- ✓ Based on ID Field
- ✓ Create View
- ✓ Preview View in Project

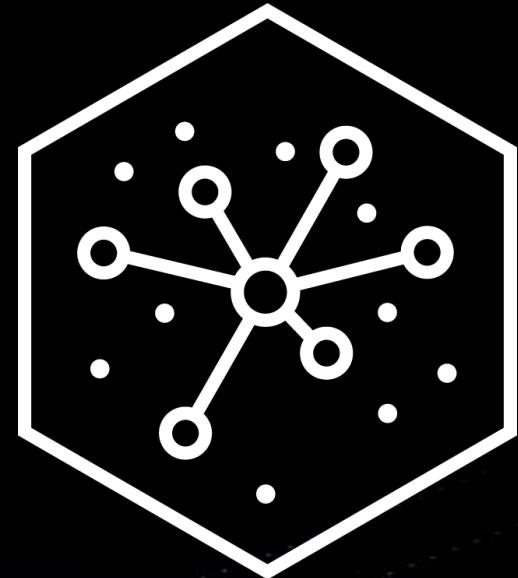
# IBM Analytics Modernization Workshop

## Part 3

<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deeper Dive</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
<ul style="list-style-type: none"><li>• <b>Analyze</b></li></ul>	<ul style="list-style-type: none"><li>• <b>Lab 06</b></li></ul>
<ul style="list-style-type: none"><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

# Analyze

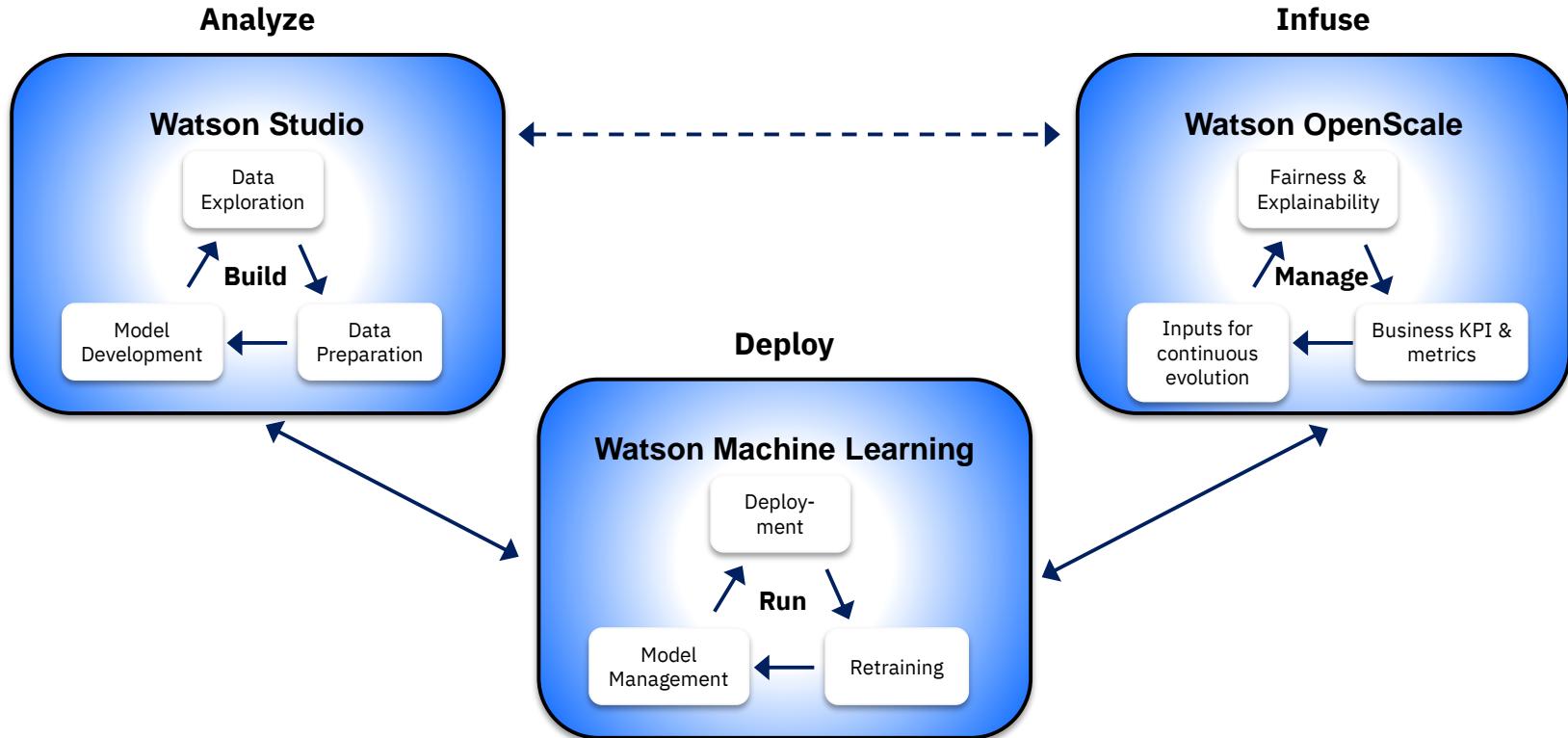
*Lab 06 – Analyze: AutoAI & Notebooks*





# Analyze

## The Data Science Lifecycle: Overview



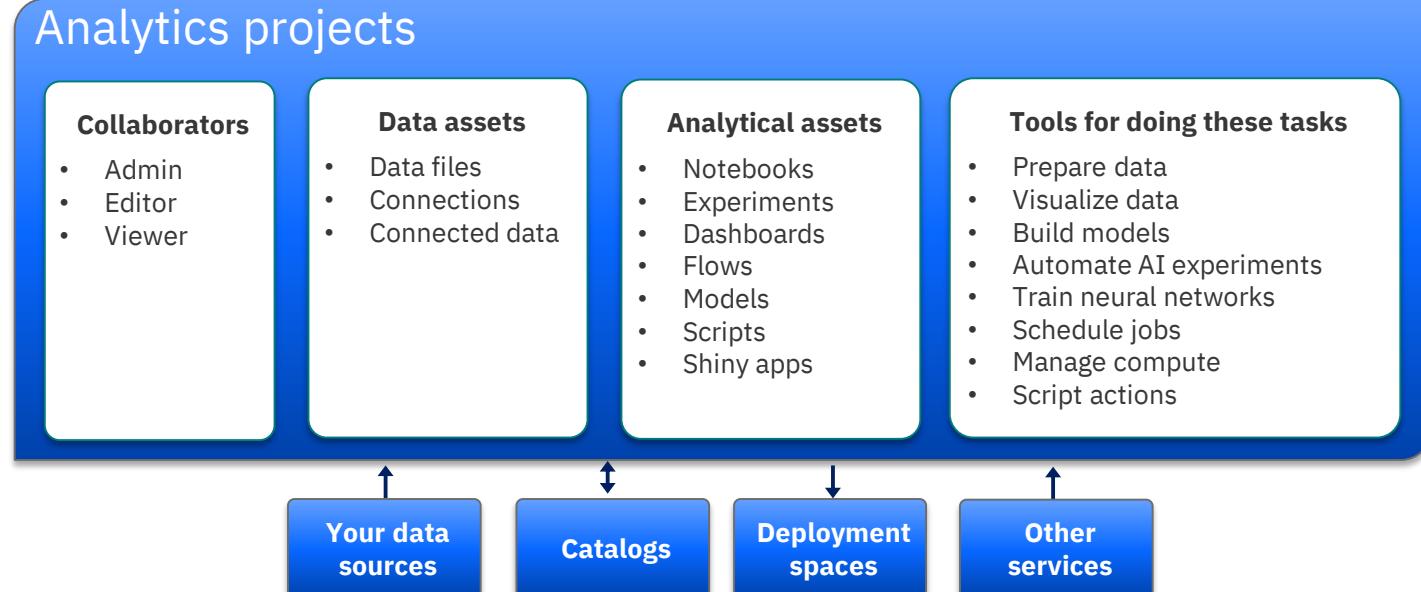


# Analyze

## Watson Studio: Collaborating with Analytics Projects

**Watson Studio** provides the environment and tools to collaborate on business problems.

**Watson Studio** is centered around the *Analytics Project*. Data scientists and business analysts use analytics projects to organize resources and analyze data with various tools.



# Analyze

## Watson Studio Integration



### Watson Studio integrates with these services:

#### Watson Knowledge Catalog

- Easily move assets between projects and catalogs
- Catalogs and projects support the same types of data assets
- Data protection rules are enforced on catalog assets that you add to projects

#### Watson Machine Learning:

- You can easily move assets between analytics projects and deployment spaces

### Watson Studio service includes these tools:

- Data Refinery
- Jupyter notebook editor
- JupyterLab IDE

### Watson Studio projects can manage these separately installed service assets:

- Watson Machine Learning AutoAI experiments
- Watson Machine Learning Accelerator DL experiments
- Cognos Dashboards Embedded
- IBM Streams flows
- SPSS Modeler flows
- Decision Optimization models
- RStudio R Shiny apps



# Analyze

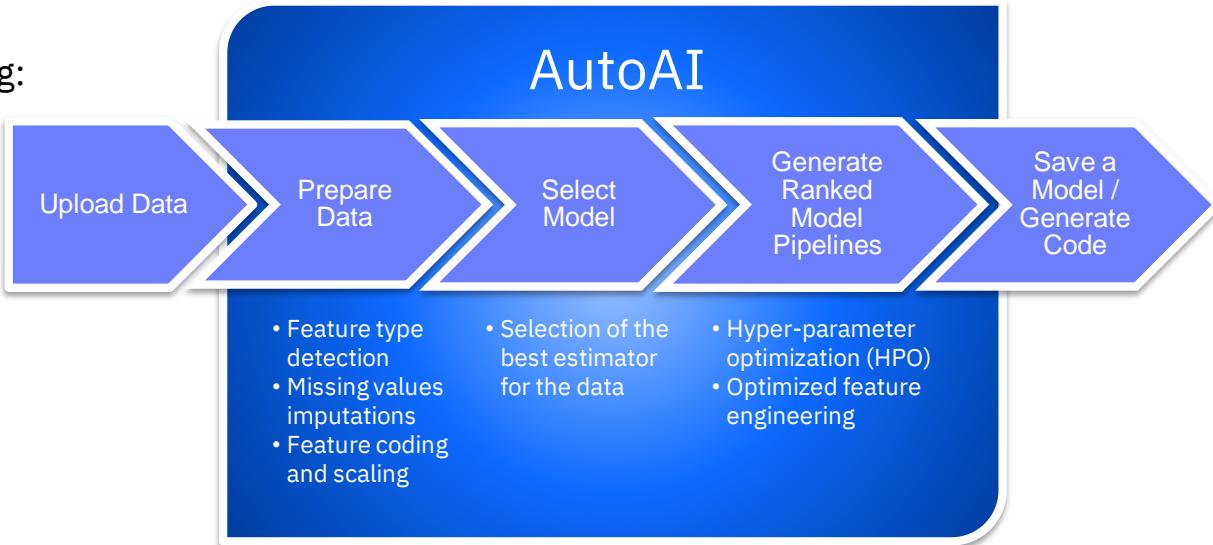
## AutoAI \* – Overview

**AutoAI** is an award-winning technology that simplifies the Machine Learning model creation and AI lifecycle by automating the following:

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

AutoAI delivers training feedback visualizations for real-time model performance results with:

- **Binary, Multiclass, and Regression support**
- **One-click model deployment**



\* AutoAI is enabled with the Watson Machine Learning service install, but it is driven through a Watson Studio Analytics Project

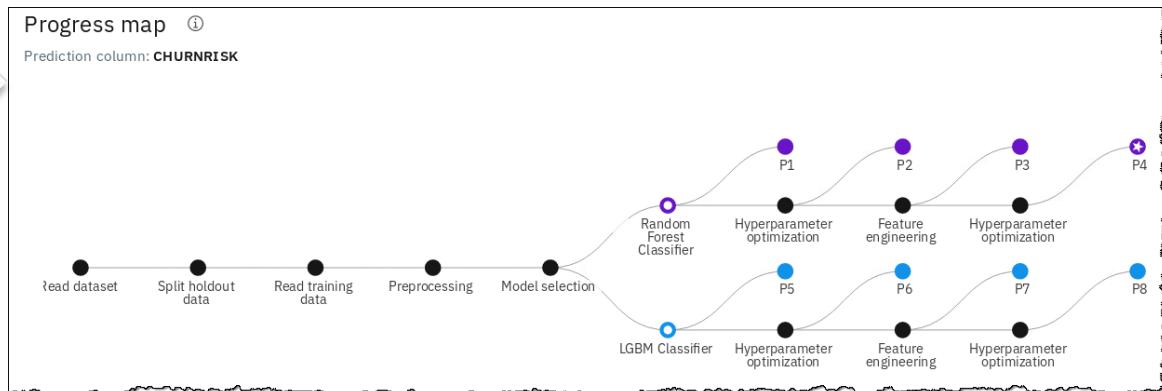


# Analyze

## AutoAI – Infographics

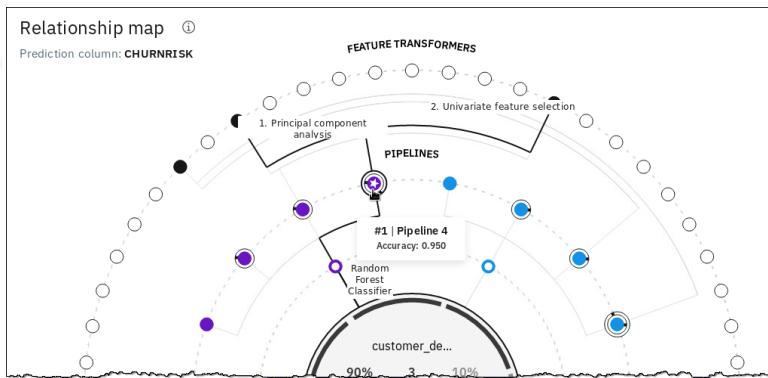
### AutoAI Progress map

Displays a progress each step as it creates the best model for your data.



### AutoAI Relationship map

Interactive infographic that shows the relationship of the pipelines, algorithms and the feature transformers.





# Analyze AutoAI – Pipelines

## AutoAI pipeline leaderboard

Shows the ranking of the pipelines for each potential model, the higher the better.

After AutoAI completes its model creation steps, you can drill into the pipeline(s) to understand how it came to its conclusion.

Save the pipeline in your project as a:

- **model**
- **notebook**

Pipeline leaderboard

Rank ↑	Name	Algorithm	Accuracy (Optimiz...)	Enhancements
★ 1	Pipeline 4	Random Forest Classifier	0.950	HPO-1 FE HPO-2
2	Pipeline 8	LGBM Classifier	0.949	HPO-1 FE HPO-2
3	Pipeline 7	LGBM Classifier	0.946	HPO-1 FE





# Analyze

## AutoAI – Benefits

### 1 Speed Model Selection

Shortlist top performing models in minutes instead of days/weeks

Drastically reduce neural network search time

### 2 Jump the Skills Gap

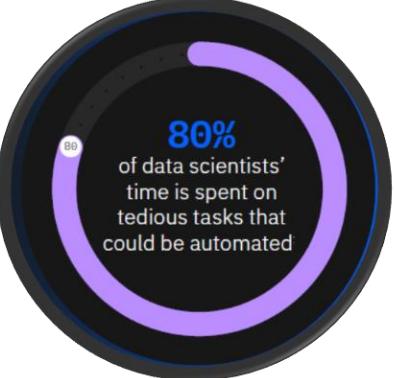
Go live with better models using the skill sets you have

Increase repeatability and minimize human intervention

### 3 Drive Productivity

Get started with AI experiments without knowing how to code

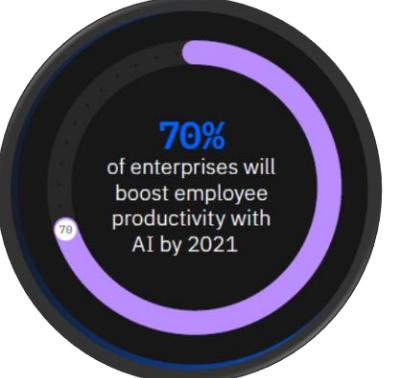
Do more innovative work instead of mundane tasks (e.g. lengthy feature selection process)



80%  
of data scientists' time is spent on tedious tasks that could be automated



63%  
of companies see availability of technical skills as a challenge to implementation



70%  
of enterprises will boost employee productivity with AI by 2021



# Analyze

## Notebooks, RStudio and other tools

**The default notebook environment:**  
Jupyter Notebook with Python 3.6

The screenshot shows a Jupyter Notebook interface with the title "TradingPlatform Customer Attrition Risk Prediction using SparkML". The notebook content discusses the use of Watson Studio Local to run analytics and predict churn. It lists three steps: 1. Ingest merged customer demographics and trading activity data, 2. Visualize merged dataset and get better understanding of data to build hypotheses for prediction, and 3. Leverage SparkML library to build classification model that predicts whether customer has propensity to churn.

**Developer tool services available:**

- Jupyter Notebooks with Python 3.6 for GPU
- Jupyter Notebooks with R 3.6
- RStudio Server with R3.6
- Lightbend Platform
- OpenSource Management



The grid displays five service cards:

- Jupyter Notebooks with Python 3.6 for GPU**: Open Source. Description: Optional development environment to create Jupyter Notebooks that use GPU-accelerated Python 3.6 libraries.
- Jupyter Notebooks with R 3.6**: Open Source. Description: Optional development environment to create Jupyter Notebooks that use R 3.6 libraries.
- Lightbend Platform**: Partner Premium. Description: Lightbend Platform makes it easy to deploy Reactive Microservices, real-time streaming and Machine Learning (ML).
- Open Source Management**: IBM. Description: Make it easy for developers and data scientists to find and access approved open source packages.
- RStudio Server with R3.6**: Partner Enabled. Description: Optional development environment for working with R.



# Analyze

## Watson Studio notebooks: Build Data Science & Machine Learning models

We split original dataset into train and test datasets. We fit the pipeline to training data and apply the trained model to transform test data and generate churn risk class prediction

```
In [67]: # instantiate a random forest classifier, take the default settings
rf=RandomForestClassifier(labelCol="label", featuresCol="features")

# Convert indexed labels back to original labels.
labelConverter = IndexToString(inputCol="prediction", outputCol="predictedLabel", labels=labelIndexer.labels)

stages += [labelIndexer, assembler, rf, labelConverter]

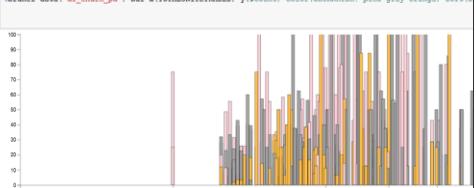
pipeline = Pipeline(stages = stages)
```

```
In [68]: # Split data into train and test datasets
train, test = df_churn.randomSplit([0.7,0.3], seed=100)
train.cache()
test.cache()
```

```
Out[68]: DataFrame[AGE: int, AGE_GROUP: string, CHILDREN: int, CHURNRISK: string, ESTINCOME: int, GENDER: string, INCOME: int, LARGESTTRANSACTIONAMOUNT: int, LARGESTTRANSACTIONCOUNT: int, MINTRANSACTIONAMOUNT: int, MINTRANSACTIONCOUNT: int, SMALLESTTRANSACTIONAMOUNT: int, SMALLESTTRANSACTIONCOUNT: int, TOTALDOLLARS: int, TOTALTRANSACTIONS: int]
```

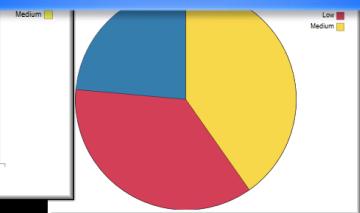
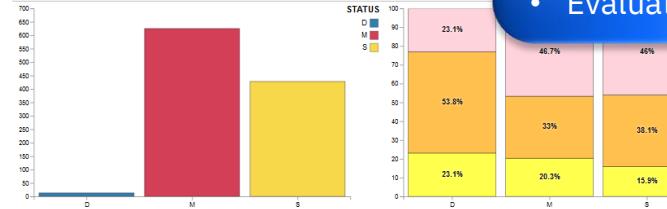


```
Out[67]:
```



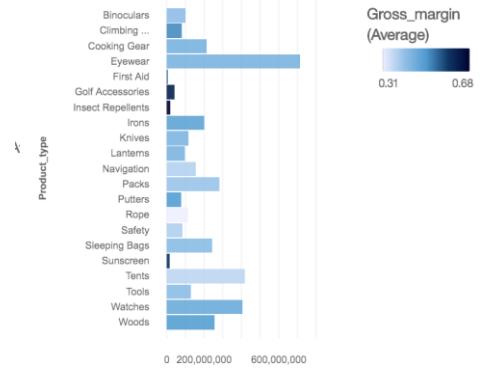
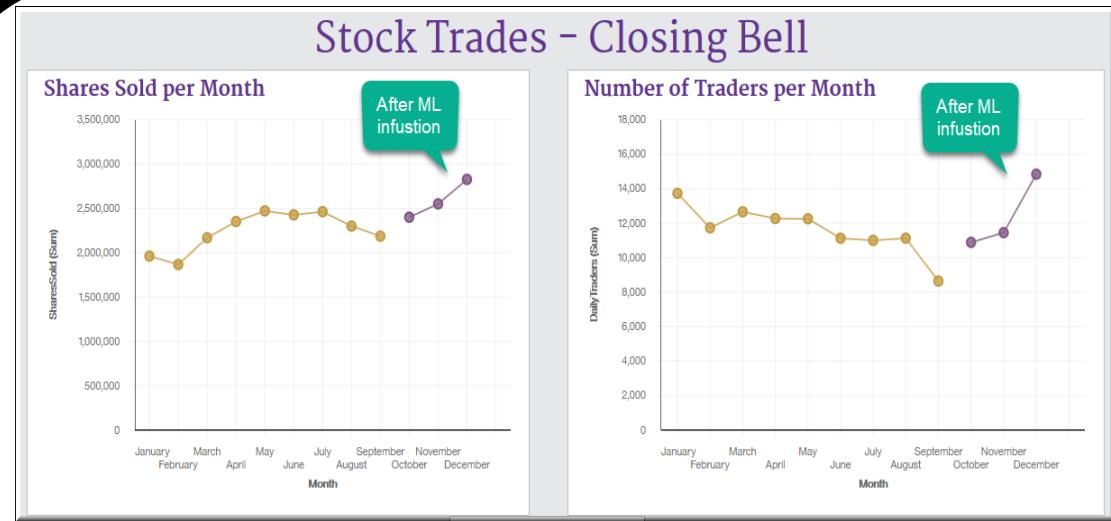
```
Out[68]:
```

- Data Scientists and Data Engineers collaborate with each other in CPD platform – while still maintaining data governance
- Collaboration using GitHub or BitBucket is integrated into the platform, which brings a cohesiveness to the work culture and helps to automate CI/CD pipe line
- Exploit GPUs for deep learning predictive ML models
- Programmatically build data visualizations and data wrangling
- Real-time or batch model scoring
- Evaluate model accuracy





# Analyze Cognos Dashboards Embedded





# Analyze IBM Streams

## IBM Streams has built-in streaming analytics

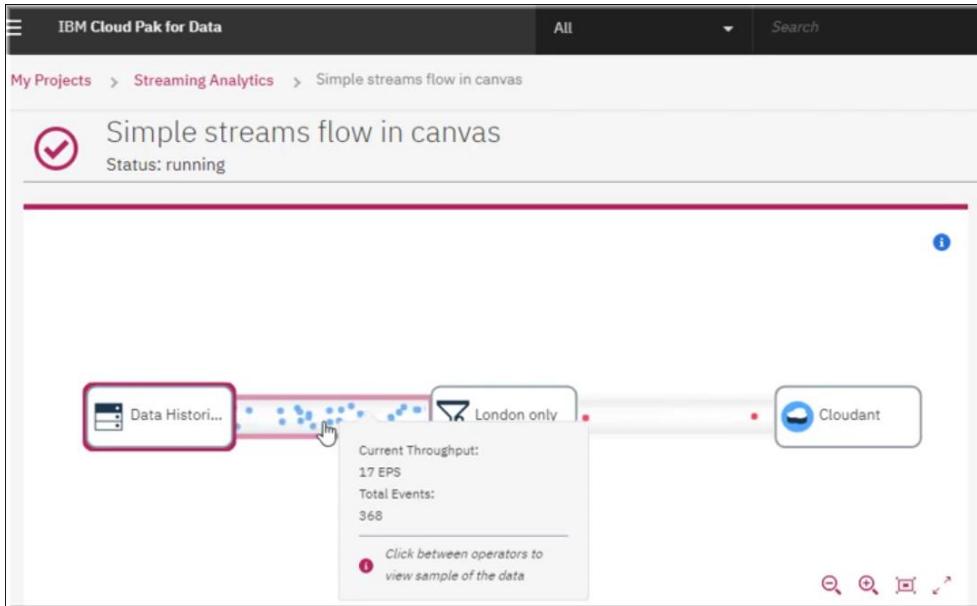
- No business disruption—run, score & update models continuously
- Machine learning, natural language, spatial-temporal, acoustic time series, etc.

## Open architecture built for speed

- Millions of events per second for massive amounts of data analytics support
- Ultra-low latency clustered runtime
- Integrate via Kafka, JSON SQL/NoSQL & more

## Rapid development

- Wizards, drag/drop development, performance dashboards, debugger
- Python, Java, Scala, PMML, R, C/C++ support
- VS Code and Atom plug-ins
- Export flows to a Python notebook



A streams flow consists of *operators*.

Every node on the streams flow canvas is an operator.

*Operator types include: sources, targets, data processing, alerts and real-time analytics*



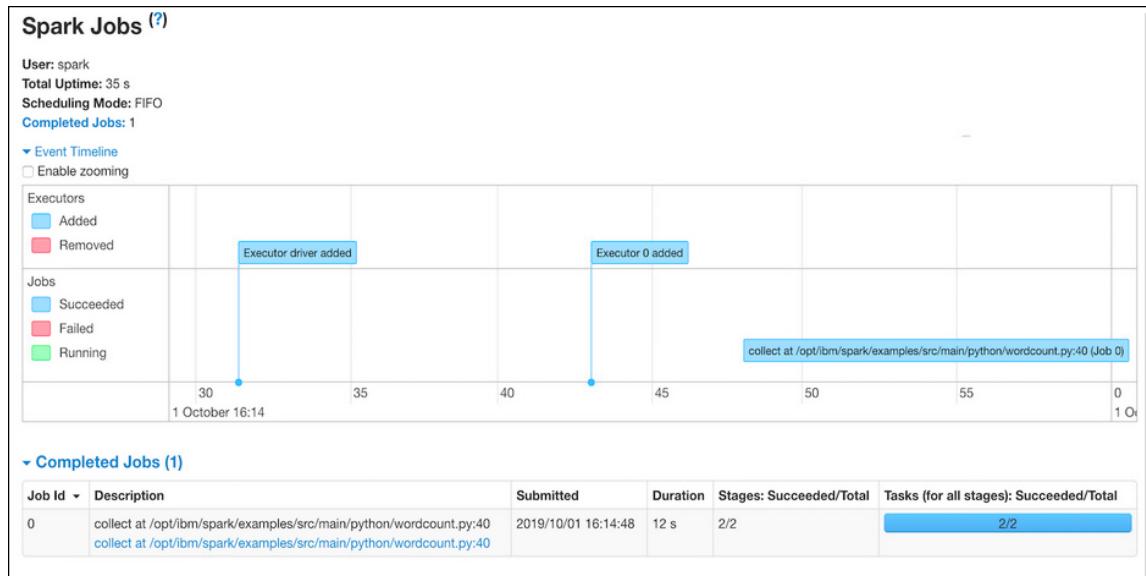
# Analyze

## Analytics Engine - powered by Apache Spark

**Analytics Engine** is a serverless, performant, customizable and dedicated Spark engine that is available in seconds.

100% open source, Analytics Engine can run a variety of workloads on the CPD cluster:

- Watson Studio notebooks that call Apache Spark APIs
- Spark application that run Spark SQL
- Data transformation jobs
- Data Science jobs
- Machine Learning jobs



# Analyze Premium Service: SPSS Modeler

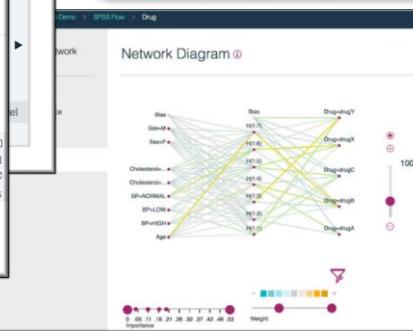


The screenshot shows the IBM Watson Premium Service interface with the SPSS Modeler workspace open. The left sidebar includes a search palette, import options, record operations, field operations, graphs, modeling (with association rules, auto classifier, auto numeric, C5.0, C&R Tree, CHAID, GLE, linear, linear-AS, and linear SVM), and a UCI ML repository. The main workspace displays a flow diagram for a Chronic Kidney Disease model. Nodes include Data Audit, Target D..., Decision ..., Partition, Decision ..., Analysis, and a central node labeled with a question mark. A 'Chart' node is connected to a histogram showing the distribution of Age. A 'Spreadsheet' node displays a table of data with columns: Age, Sex, BP, Cholesterol, Na, and K. The table rows show values for 10 different individuals.

	Age	Sex	BP	Cholesterol	Na	K
1	23	F	HIGH	HIGH	0.739535	0.031268
2	47	M	LOW	HIGH	0.739309	0.056466
3	47	M	LOW	HIGH	0.697269	0.068944
4	28	F	NORMAL	HIGH	0.563682	0.072289
5	61	F	LOW	HIGH	0.559294	0.030968
6	22	F	NORMAL	HIGH	0.678901	0.078647
7	49	F	NORMAL	HIGH	0.789637	0.048598
8	41	M	LOW	HIGH	0.766535	0.069461
9	60	M	NORMAL	HIGH	0.777205	0.05123
10	43	M	LOW	NORMAL	0.526102	0.027164

## SPSS Modeler

- A leading visual data science and machine-learning and predictive analytics solution
- Helps enterprises accelerate time to value and achieve desired outcomes by speeding up operational tasks for data scientists and business analysts
- Tap into data assets and modern applications, with complete algorithms and models that are ready for immediate use



# Analyze

## Premium Service: Decision Optimization



**Decision Optimization (DO)** enables data science teams to capitalize on the power of *prescriptive analytics* and build solutions using a combination of techniques like optimization and machine learning.

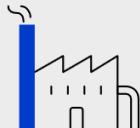
Integrated with Watson Studio, Decision Optimization can combine optimization techniques with coding and non-coding tools, model management and deployment – as well as other data science capabilities.

Decision Optimization evaluates millions of possibilities – balancing trade-offs and business constraints to find the best possible solution.

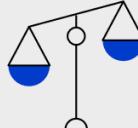
Insights that drive optimal decisions to complex problems



Determine location  
and capacity  
of warehouses



Determine which plant  
should manufacture  
which product



Build financial  
portfolios by balancing  
risks and rewards



Allocate aircraft  
and crew to flights



# Analyze

## Watson Machine Learning : WML Accelerator for Deep Learning

The **Experiment Builder** GUI interface is available from a WML project when you install the **WML Accelerator**. It is the simplest method to perform Deep Learning experiments.

The Watson Machine Learning Accelerator has the following components:

- **IBM Spectrum Conductor Deep Learning Impact version 2.1.0:**  
Provides robust, end-to-end workflow support for deep learning application logic for the complete lifecycle management from data ingest and preparation to building, optimizing, training and testing the model.
- **IBM Spectrum Conductor version 2.4.0:**  
A *highly available* and resilient *multitenant distributed* framework, providing deep learning application lifecycle support, centralized management and monitoring, and end-to-end security.
- **Deep Learning Frameworks:**  
TensorFlow, PyTorch, Keras, and Caffe
- **Deep Learning Rest API**



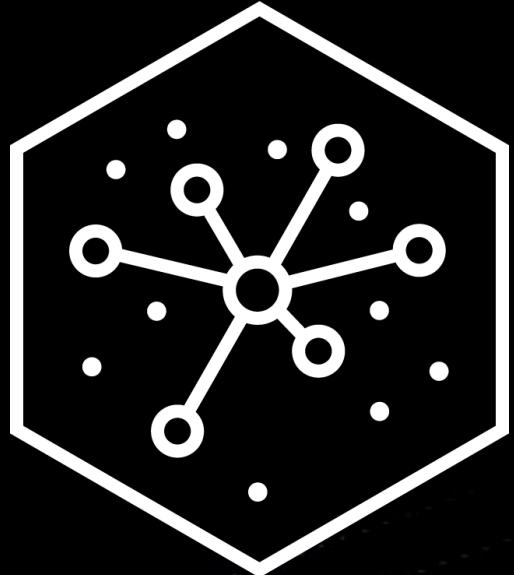
### Key Features

- Supported AI frameworks with performance optimizations for GPU acceleration
- Transparent GPU Topology for Distributed Training
- Auto Hyper Parameter Optimization (HPO)
- Multi-Tenancy for Training and Inference
- Elastic Distributed Training (EDT)
- Resource Utilization, Monitoring, and Reporting
- Elastic Distributed Inference (EDI)
- Secure Deployment Model
- Role Based Access Control / Kerberos Authentication

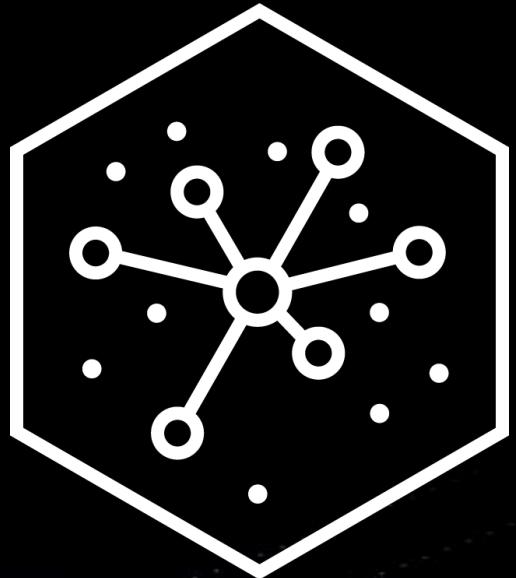
# Cloud Pak for Data 3.0.1

Work on Lab 6

Session restarts at xx PM

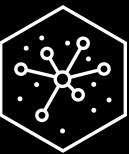


# Wrap up



# Cloud Pak for Data

## Unified Experience

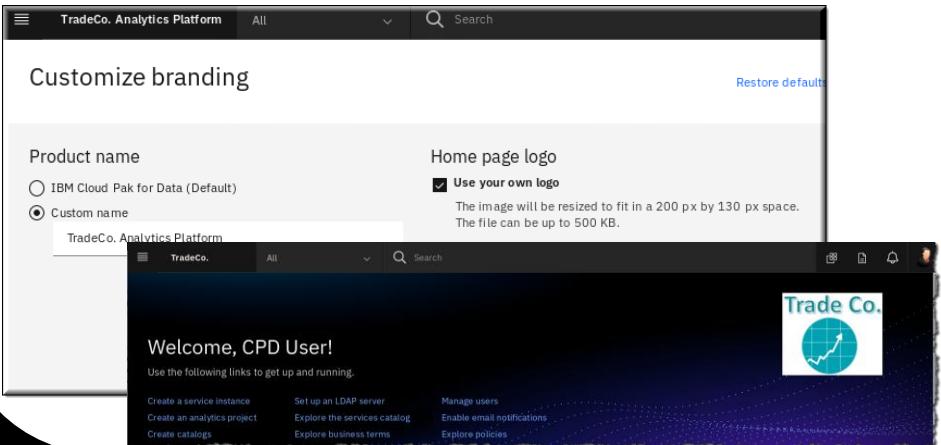


### Customized Logo and Branding

Example: Customize the web client interface per tenant.

Customizable components in the Home Page:

- Product name
- Home page logo



### Group 1 language support

All services in Cloud Pak for Data are translated for the following languages:

- Simplified Chinese
- Traditional Chinese
- Japanese
- French
- German
- Italian
- Spanish
- Brazilian Portuguese

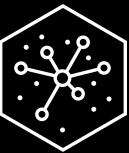
### Carbon 10

Modernized, modular and flexible open source design framework

- Consistent look and feel
- Reusable components

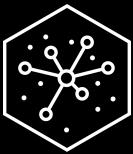
# Cloud Pak for Data

## Version 3.0.1 notable updates



Update	Comment
Support for POWER systems	Build an AI foundation for speed and scale with the premier, built-in GPU acceleration platform for faster time to AI.
OpenShift Container Storage	Software-defined Storage automated for quicker and efficient hybrid multi-cloud deployments and optimized for Red Hat OpenShift Platform.
Red Hat OpenShift 4.x support	Or continue to run on RHOC 3.11.
IAM Integration	IBM Cloud Platform Common Services (CPCS) IAM for all Cloud Paks provides authentication support via the OpenID Connect (OIDC) specification for SSO capability.
Fine-grained permissions	4 New fine-grained permissions: <ul style="list-style-type: none"><li>Configure authentication</li><li>Configure platform</li><li>Manage users</li><li>Monitor platform</li></ul>
Operations Management	New utilities for upgrade, backup and restore, import and export

# Cloud Pak for Data Security Considerations



## Security Features

- ✓ Security Architecture and Design
- ✓ Access Control, Authentication and Authorization (e.g. integrates with leading LDAPs)
- ✓ Data Protection
- ✓ Security Logging

## Security Engineering

- ✓ Development trained in Secure Coding Practices
- ✓ Secure Engineering Development Practices: threat modeling, risk assessment, static and dynamic code analysis, penetration testing, container scanning, etc.

## Security Operations

- ✓ Audit Log consolidation and analysis
- ✓ User access management
- ✓ Security Incident Management

## Governance & Compliance

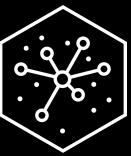
- ✓ Compliance Controls defined by Outside Agencies
- ✓ System Security Plans for maintaining compliance security postures

## Compliance Best Practices:

**FISMA** High “ready” with System Security Plans, spanning 350 controls:

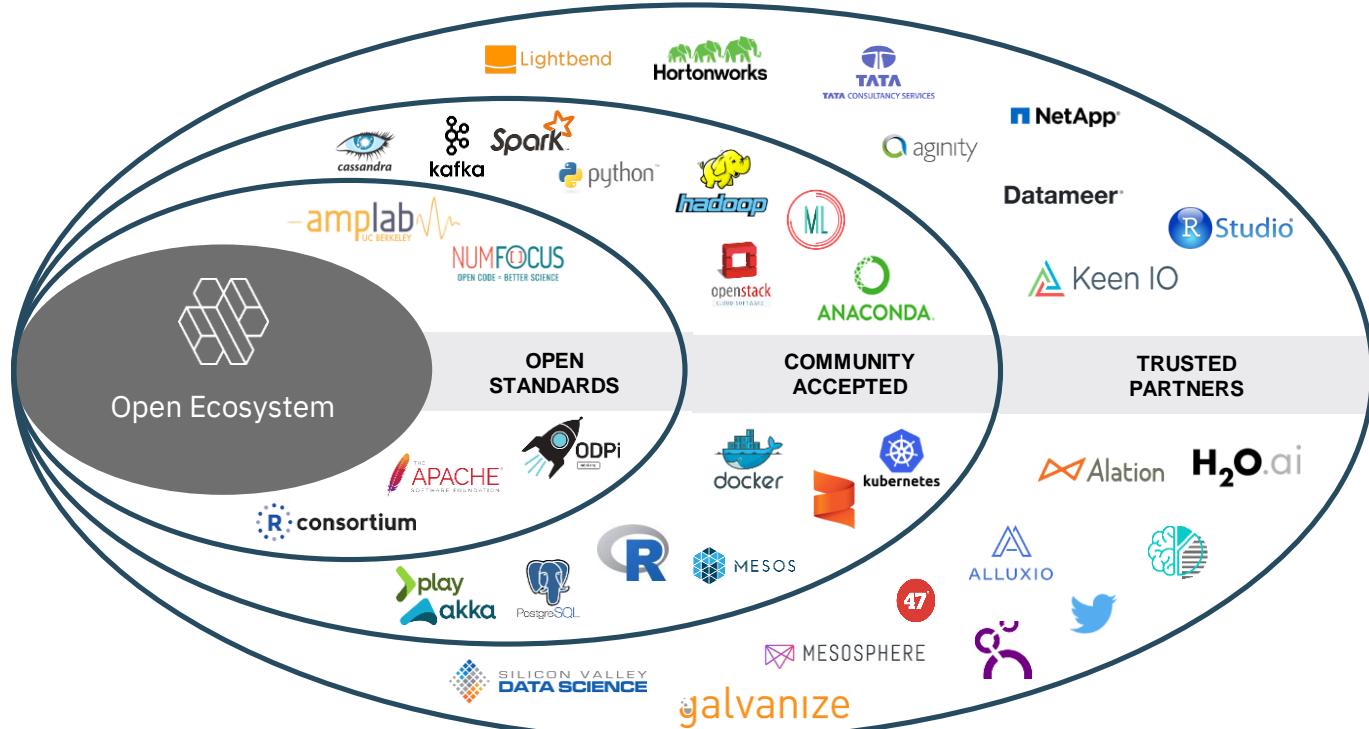
- Risk Assessment
- Certification, Accreditation and Security Assessments
- System Services and Acquisition
- Security Planning
- Configuration Management
- System and Communications Protection
- Personnel Security
- Awareness and Training
- Physical and Environmental Protection
- Media Protection
- Contingency Planning
- System and Information Integrity
- Incident Response
- Identification and Authentication
- Access Control, Accountability and Audit

**GDPR** “readiness” considerations



# CPD built on an open Ecosystem

## Where IBM leads, partners and co-creates



IBM's approach to Open technology: <https://developer.ibm.com/articles/cl-open-architecture-update/>



Build Once - Run anywhere  
In your own data center  
Or the cloud infrastructure of your choice



Your Data Center



IBM Cloud



OPENSHIFT



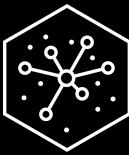
openstack™



*Helps avoid vendor lock in*

# CPD Use Cases

## Industry agnostic



### Create a Customer Focused Enterprise

- Rich profile for every customer kept up to date in real time as new customer behaviour is collected (360 view)
  - Spending Patterns
  - Behaviour
- Deliver tailored offerings based on segmentation
- Provide “next best action” in real-time

***These use cases apply to most industry verticals***

# CPD Use Cases

## Industry specific

Over 24 Data Science & AI use-cases across 15 industry verticals (*applicability varies by customer*)

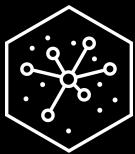
### Cloud Pak for Data differentiation :

- Operationalize models in a matter of minutes (Deploy, scale & manage models with minimum effort)
- Model governance (Lineage and provenance – Who created, when, what data was used, comments, ratings etc.)

Use Case(s)	Aerospace & Defense	Automotive	Banking	Chemicals & Petroleum	Consumer	Education	Electronics	Energy, Environmental & Utilities	Financial Markets	Government	Healthcare & Life Sciences	Insurance	Industrial Products	Telco, Media & Entertainment	Travel & Transportation
Predictive Maintenance	X							X							X
Real time analytics (IOT)	X	X		X											X
Customer Churn / Retention		X	X												X
Anomaly Detection	X	X						X						X	
Regulatory Compliance				X					X			X			
Anti-Money Laundering (AML)				X											
Cross Sell / Up-Sell			X		X										
Demand Forecasting				X			X	X							
Inventory Optimization					X										
Retention & Time to Degree								X							
Application modernization								X							
Student Safety							X								
Predictive Customer Insights				X				X							
Counter Fraud & Payments									X			X			
Counter Party Credit-risk											X				
Client Insights for Wealth Management											X				
Threat Prediction & Prevention											X				
Patient Diagnosis												X			
Data Privacy												X			
Client Risk Scoring			X										X		
Targeted Ads														X	
Intrusion Detection	X													X	
Route Optimization															X

# Cloud Pak for Data

## Ranked #1 by Forrester for “Enterprise Insight Platforms”



### Enterprise Insight Platforms - Definition

- Enterprise insight platforms pre-integrate most — or all — of the technology required to build systems of insight and thus help business move faster. The need to move faster and change more easily is the driving force behind customer demand for these platforms.
- Vendors that can better support all the personas of an insight team with unified experiences that feature governance and can creatively enable hybrid cloud and multi-cloud delivery will win.

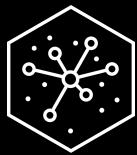
### Forrester on “ICP for Data”

IBM has an impressive portfolio of individual data management and analytics capabilities that have consistently scored well on individual component Forrester Waves. With ICP for Data, IBM has pre-integrated capabilities that allow clients to be productive in a week or less. We were also impressed with its ML-assisted data cataloging and governance tools. IBM's platform uses Kubernetes to deploy on-premises or into the public cloud. Lastly, IBM's support for different insight team personas through tailored but unified experiences is commendable. Firms looking to unify the work of insight teams will do well on this platform.

### Forrester on Microsoft’s Perceived Weakness – Azure Cloud Platform

While Microsoft offers AI services, its multimodal predictive analytics and machine learning (PAML) tools scored poorly in previous Forrester Waves. Finally, we found this offering to be too light on data governance capabilities and self-service data preparation tooling, both of which are critical insight team capabilities.

Report Preview : <https://ibm.box.com/s/bry68nm9alduszvrffo105cvjhmd7pn>



## Cloud Pak for Data Editions

Make your data ready for AI – Cloud Agility, Lightning Fast & AI-ready

### Standard Edition

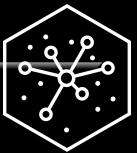
- Beginning with a minimum of 24 VPCs
- Expands in increments of 1 VPC
- Up to a *maximum* of 64 VPCs
- You cannot use separately priced premium services with this edition
- 50% Enterprise Edition list price

### Enterprise Edition

- Deploy in any form: on premises, on public cloud or CPD System
- Beginning w/ recommended minimum of 48 VPCs
- Select 24 VPC configurations supported
- Expands in increments of 1 VPC
- No maximum

### Non-Prod Edition

- Can only be used for non-production scenarios
- No restriction in terms capabilities or VPC quantities
- Needs to be in separate namespace from production licenses
- Need parallel and appropriately sized Standard or Enterprise Edition licenses for production use
- 50% of Enterprise Edition list price



# IBM Cloud Pak for Data System

True plug-and-play enterprise data & AI in hours right out of the box



- Brings the elasticity & scalability of public clouds securely behind the firewall

- Connect all your data for self-service analytics
- Operationalize AI with trust & transparency
- Deploy dynamic cloud native data workloads

An **all-in-one** data & AI system with all the necessary systems and software components pre-integrated

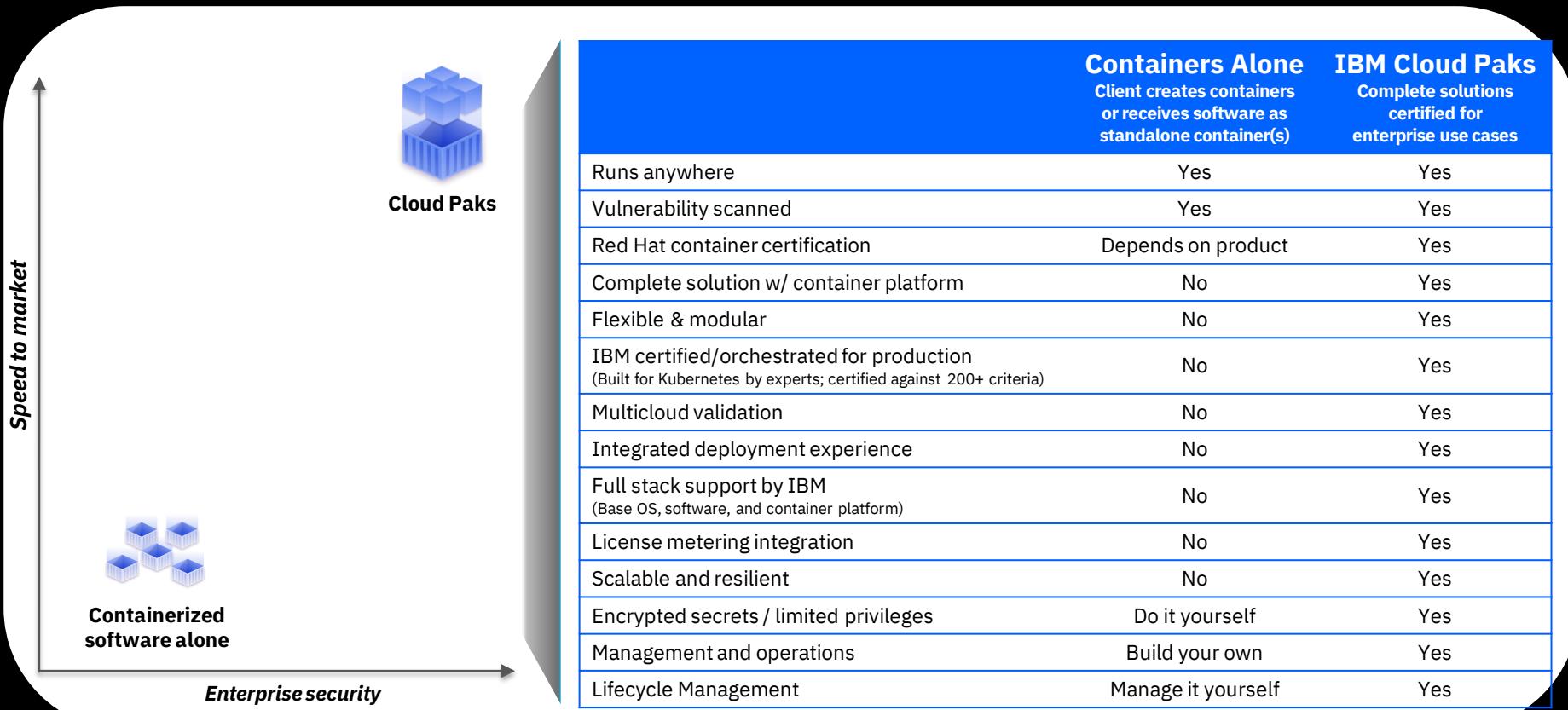
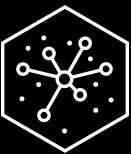
Deploy a complete private cloud in under 4 hours, with no assembly required

Dynamically scale compute, storage and networking resources with plug and play of new hardware nodes

Simplify management and optimization with a unified and intuitive dashboard

# Cloud Pak for Data vs. Containers Alone

## IBM Certified and production ready



# Cloud Pak for Data

## IBM Kubernetes Certified

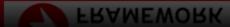
<http://ibm.biz/cp-certify>



Production Grade	Security	Quality Assurance	Lifecycle Management
 A blue circular icon containing a white key symbol, representing security features.	 A green circular icon containing a white padlock symbol, representing security management.	 A purple circular icon containing two interlocking gears, representing quality assurance processes.	 A yellow circular icon containing a gear and two blue arrows pointing in opposite directions, representing lifecycle management.

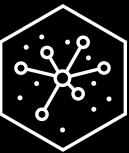
### Consistency and Standards

 A black circular icon containing a white checklist with a checkmark, representing consistency and standards.	<ul style="list-style-type: none"><li>• Consistent Packaging / Publishing</li><li>• Supporting Operators and Helm</li><li>• Consistent Entitlement management</li><li>• Common management of OSS elements</li></ul>	<ul style="list-style-type: none"><li>• UBI and Red Hat Certified</li><li>• Consistent use of OCP and IBM Services</li><li>• ~200 Code Standards enforced</li><li>• Governed Best Practice / Anti Practices</li></ul>
--	---	---

 The Red Hat OpenShift logo, featuring the Red Hat circle and the text "Red Hat OpenShift".	 The Operator Framework logo, featuring a red lightning bolt icon and the text "OPERATOR FRAMEWORK".	 The Red Hat Certified logo, featuring a red hat icon and the text "redhat CERTIFIED".	<ul style="list-style-type: none"><li>• Managed Image CVEs</li><li>• Packaging</li><li>• Publishing</li></ul>	<ul style="list-style-type: none"><li>• Trusted Source</li><li>• E2E Support</li></ul>
 The IBM Cloud Pak for Data logo, featuring the IBM arrow and the text "IBM CLOUD PAK FOR DATA".	 The IBM Cloud Pak for Data logo, featuring the IBM arrow and the text "IBM CLOUD PAK FOR DATA".	 The IBM Cloud Pak for Data logo, featuring the IBM arrow and the text "IBM CLOUD PAK FOR DATA".	<ul style="list-style-type: none"><li>• Upgrades</li><li>• Backups</li><li>• Restores</li></ul>	<ul style="list-style-type: none"><li>• E2E Support</li><li>• Monitoring</li><li>• Logging</li></ul>

# Cloud Pak for Data

## Key differentiators (vs. Microsoft, AWS and Google)



### Data virtualization

- Query all of your data sources as one
- Governance, security, and scalability by design
- 400X faster than federation
- *Unmatched by Microsoft, AWS, or Google (may only have simple federation at best)*

### Data governance

- Data privacy & governance by design: data discovery & curation, with policy & rules management
- Metadata management and shopping for data
- Smarter compliance: Regulatory ML, industry accelerators, FISMA HIGH certification, etc.
- *Unmatched by Microsoft, AWS, or Google*

### Model bias and drift detections w/ explainability

- Detect and mitigate model bias and drift
- Explainability of a model prediction for a single transaction
- *Unmatched by Microsoft, AWS, or Google (none have all three)*

### Governing and operationalizing AI

- Governed AI lifecycle management
- CI/CD style pipelines for AI DevOps
- AI model trust and transparency
- *Unmatched by Microsoft, AWS, or Google*

### Open source based hyperconverged system

- Query all of your data sources as one
- Governance, security, and scalability by design
- 40X faster than federation
- *Unmatched by Microsoft, AWS, or Google*



# Turbocharged digital transformation

Read the story in any of these magazines:

## Business Chief US

(front cover, story pages 12-23,  
Cloud Pak for Data pages 16-17)

## Business Chief Canada

(story pages 144-153, Cloud Pak for Data pages 146-147)

## Gigabit Magazine

(front cover, story pages 12-23, Cloud Pak for Data pages 16-17)



Facing its own path toward digital transformation, Sprint started preparing its data for artificial intelligence (AI) with the goal of using machine learning algorithms to gain quicker insights and increase responsiveness to customers.

Sprint chose [Cloud Pak for Data](#) because it enables AI projects in weeks rather than months through unifying and simplifying three critical stages in the journey to AI: the collection, organization and analysis of data.

## Cloud Pak for Data

Industry: Telecommunications  
Geography: North America

Simplify and automate how your organization turns data into insights within a unified, all-in-one design.

*"Cloud Pak for Data enabled Sprint to digest high volumes of data for near real-time ML/AI analysis, and the trial results have shown potential to take Sprint to the next phase of digital transformation."*

Michelle Gehl

VP Networks OSS Applications and Operations, Sprint





# Is AI your priority? Start with a data strategy

[Read the blog and  
watch the video](#)

Intel and IBM are great partners and closely aligned in becoming more data-centric.

Intel's participation and contribution is meaningful because customers can run [Cloud Pak for Data](#) at speed on their Intel-based infrastructure. The union between IBM and Intel is supercharging the ability of data scientists to drive better insight and better business outcomes in a way that has never been seen before.

## Cloud Pak for Data

Industry: Technology

Geography: North America

Simplify and automate how your organization turns data into insights within a unified, all-in-one design.

*"Cloud Pak for Data is really important because it helps to do a couple of things that are mind blowing for data scientists — auto discovery of data and rapid integration of hyper-relevant data."*

Melvin Greer

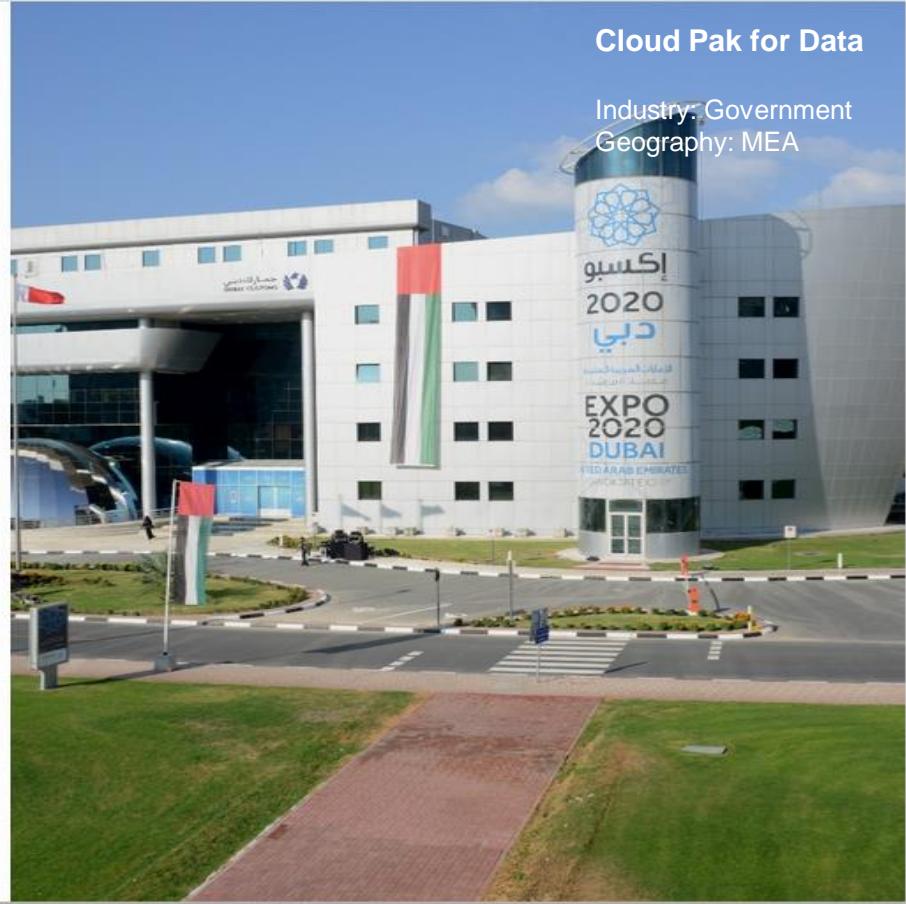
Senior Principal Engineer and Chief Data Scientist  
- Americas, Intel Corporation

**Government | MEA (United Arab Emirates)**

# Accelerating the Customs Process

With Cloud Pak for Data's end-to-end platform, Dubai Customs is building trusted AI models to help identify risky goods entering the country and reduce the number of false positives.

- **When:** Current | **Duration:** 3 months
- **Product:** Cloud Pak for Data, Watson OpenScale
- **Pillars:** DSAI



**Cloud Pak for Data**

Industry: Government  
Geography: MEA

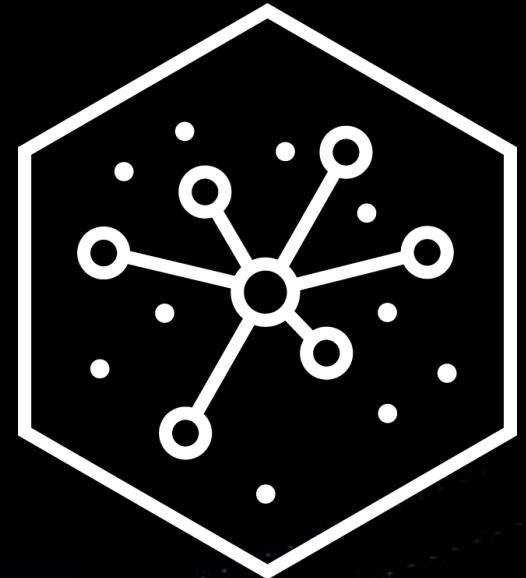
# IBM Analytics Modernization Workshop

## Part 3

<ul style="list-style-type: none"><li>• Introduction</li><li>• Business Use Case</li></ul>	<ul style="list-style-type: none"><li>• Lab 01</li><li>• Lab 02</li></ul>
<ul style="list-style-type: none"><li>• Collect: Connect</li><li>• Organize – Deeper Dive</li><li>• Collect: Virtualize</li></ul>	<ul style="list-style-type: none"><li>• Lab 03</li><li>• Lab 13</li><li>• Lab 05</li></ul>
<ul style="list-style-type: none"><li>• <b>Analyze</b></li></ul>	<ul style="list-style-type: none"><li>• <b>Lab 06</b></li></ul>
<ul style="list-style-type: none"><li>• Deploy</li><li>• Infuse – OpenScale</li><li>• Infuse – Cognos Analytics</li><li>• Wrap up</li></ul>	<ul style="list-style-type: none"><li>• Lab 07</li><li>• Lab 08</li><li>• Lab 09</li><li>• Lab 10</li></ul>

**Thank you for your time!**

**Begin your journey now on the  
IBM Platform built for AI...**



**We appreciate your feedback.**

# Copyright and trademarks

© Copyright IBM Corporation 2020

IBM Corporation  
Route 100  
Somers, NY 10589

Produced in the United States of America  
July 2020

IBM, the IBM logo, ibm.com, API Connect, Db2, Elastic Storage, FlashCore, POWER, Spectrum Scale, UrbanCode, WebSphere and IBM Z are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)

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.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both. Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.