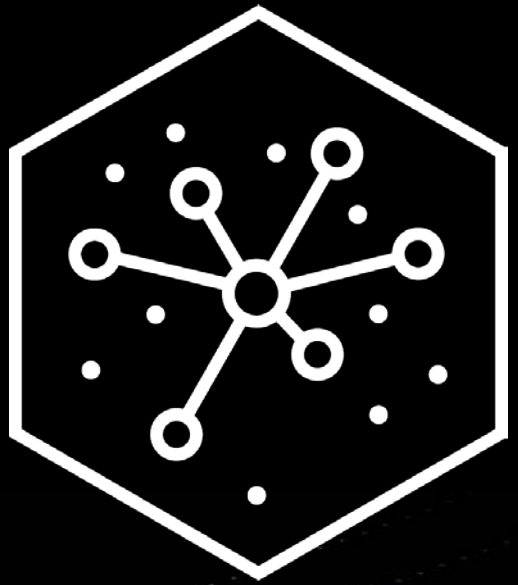
IBM Journey to Cloud and AI Analytics Modernization Workshop

Featuring: Cloud Pak for Data 3.0.1



Presentation by: Burt Vialpando and Kent Rubin - IBM

October 28, 2020

Welcome to the IBM® Briefing Center

Location logistics

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

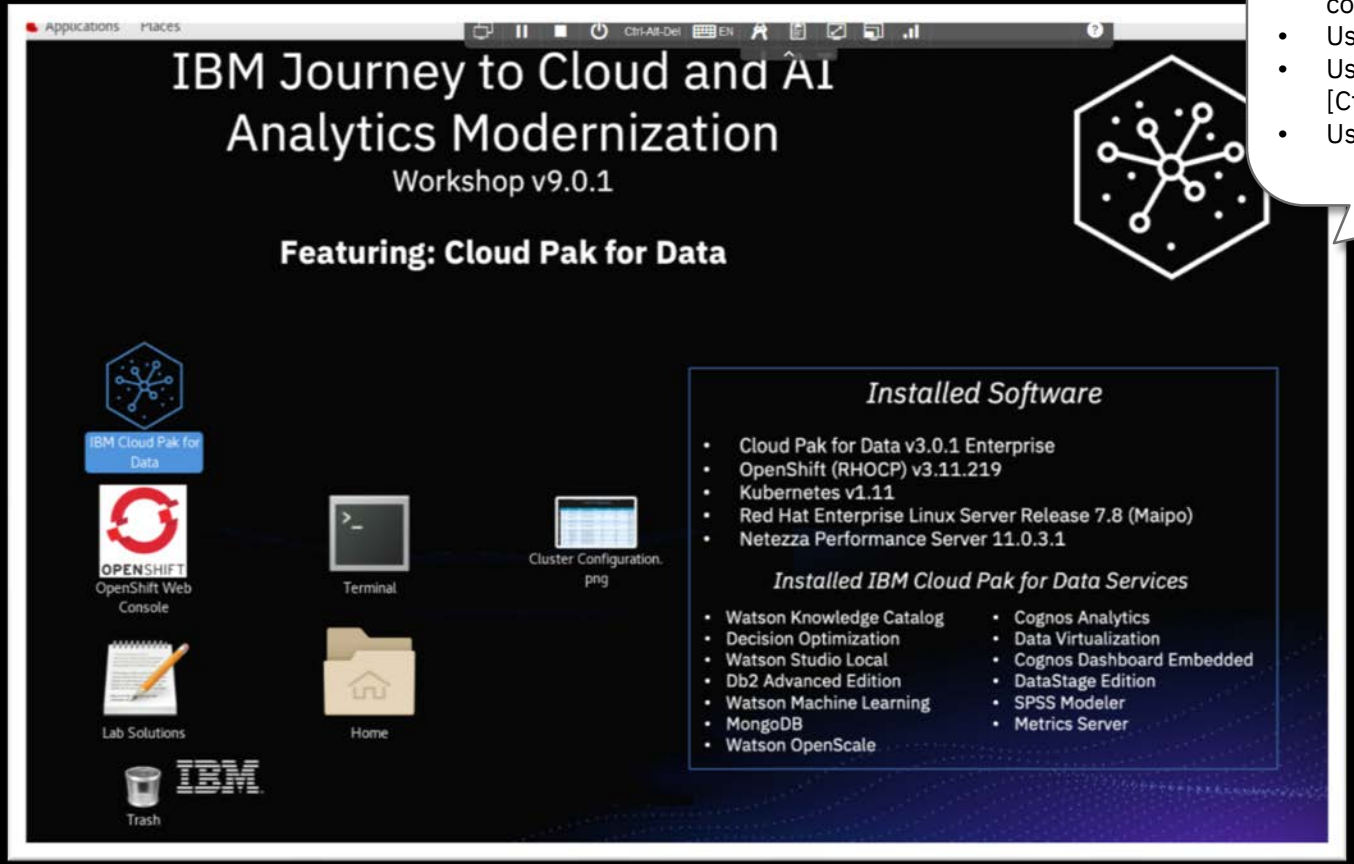
Introductions

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

IBM Analytics Modernization Workshop
Agenda

Part 1	• Introduction • Business Use Case	• Lab 01 • Lab 02
Part 2	• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
Part 3	• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

IBM Analytics Modernization Workshop
The workshop Unified Desktop



Workshop lab exercise tips

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

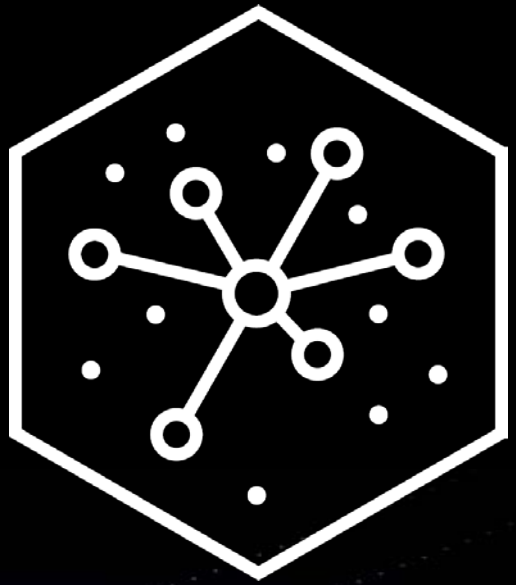
IBM Analytics Modernization Workshop

Part 1

• Introduction • Business Use Case	• Lab 01 • Lab 02
• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

Introduction

Lab 01 – Getting Started

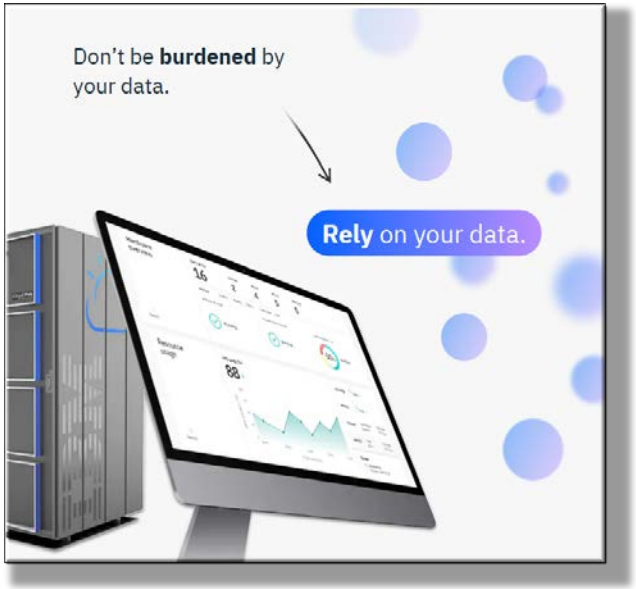


IBM Cloud Pak for Data

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.



Why Cloud Pak for Data?
The business case

The Challenge

Enterprises are not leveraging the data that is created

- ***2.5 quintillion bytes*** *of data is created around the world EVERY DAY!*
- ***~70% of a company's data remains unused****, yet every company must be data driven in today's ecosystem*
- ***Only 7.3%*** *of leading company executives are confident in their own strategy for leveraging data*

The Factors

Most enterprises are challenged to be truly data driven

- ***Most large enterprises*** *have legacy systems making it difficult to connect to and utilize all their data*
- ***Most industries*** *have to deal with increasing regulatory constraints to protect their data*
- *Highly paid teams spend* ***50-80%*** *of their time to find, prepare, and govern data assets with increasing silos*
- ***97.2%*** *of leading enterprise executives launch data and AI initiatives due to the knowledge that three-quarters of Fortune 1,000 companies have been replaced by data driven competitors*

The Solution

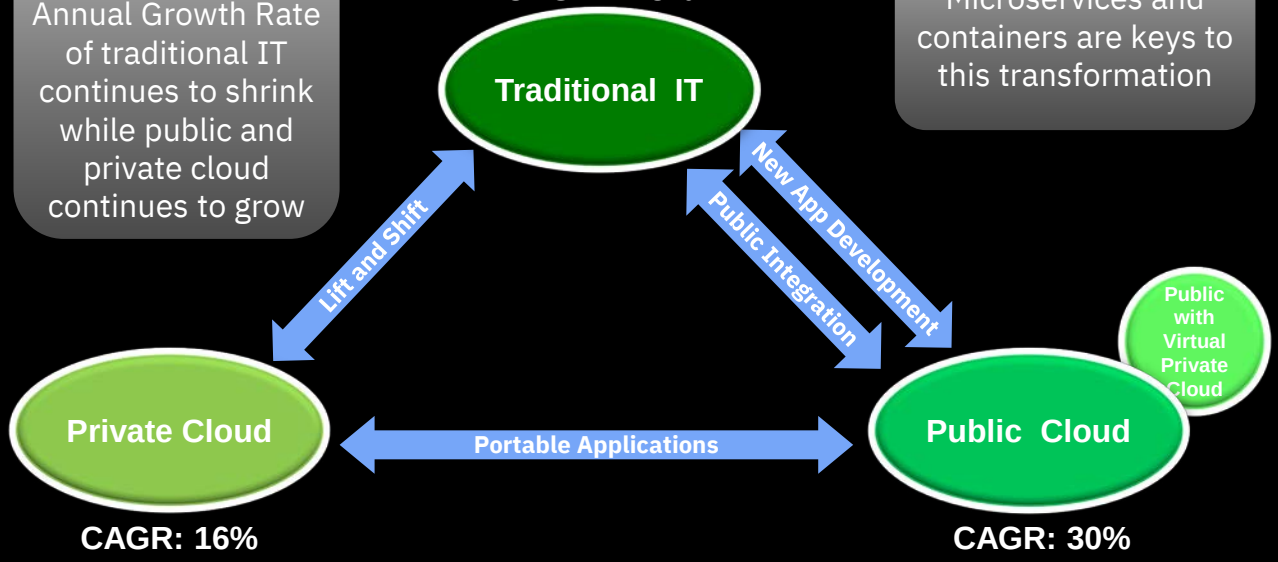
Cloud Pak for Data!

- *Comprehensive, integrated, and extensible platform to make data ready for AI*
- *Built on enterprise grade* ***private cloud*** *that is reliable, efficient, scalable and portable*
- ***The right technology for the right job!***

Cloud Pak for Data: built on cloud-native architecture

Multi-cloud is being driven by cloud-native architectures
Microservices and containers are changing IT

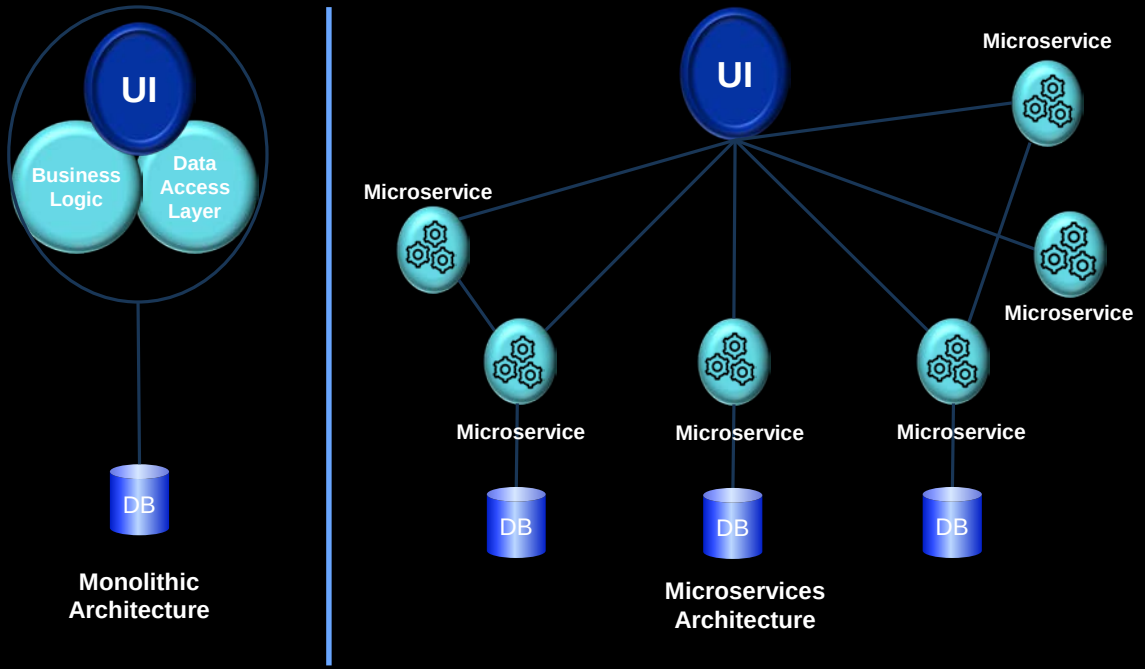
The Compound Annual Growth Rate of traditional IT continues to shrink while public and private cloud continues to grow



Microservices and containers are keys to this transformation

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:***

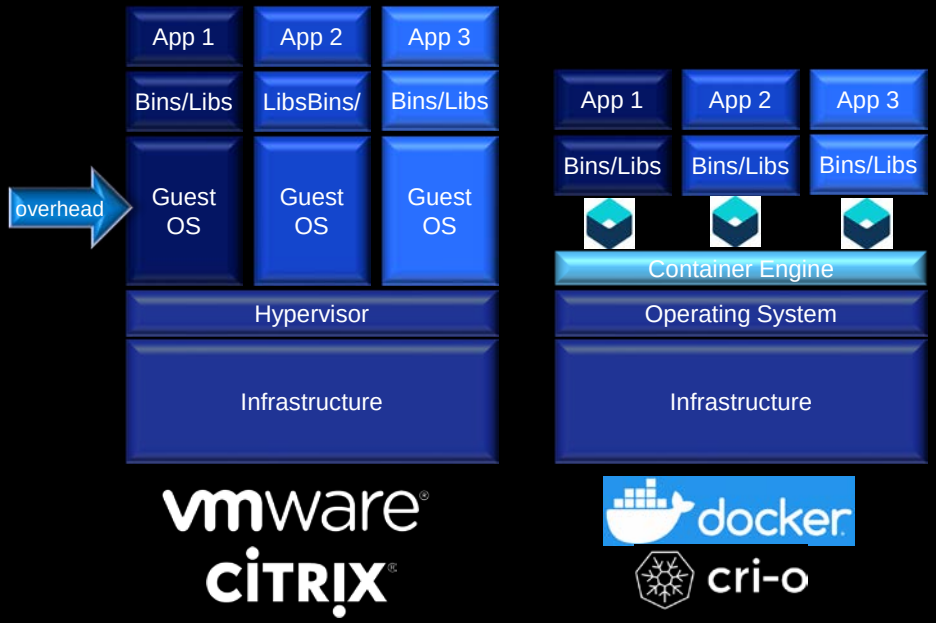
  Auto provision, auto scale and provide auto-redundancy

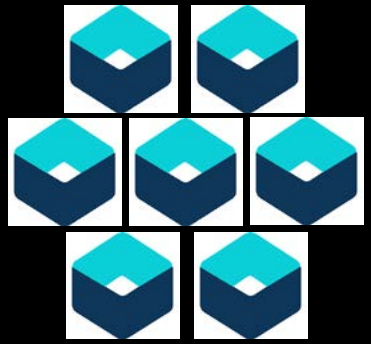
* This is not a claim that a microservice-based application approach is always better for every use case scenario

Containers – the second key to cloud native applications

Reducing operational and development costs

Virtual machines vs. containers *





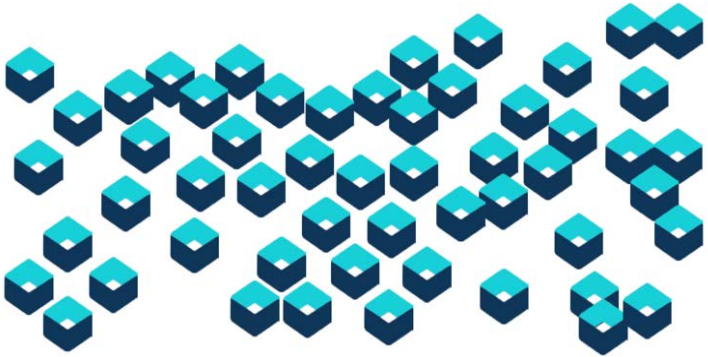
Containers can be 2 – 3 times more resource efficient than virtual machines

On average Docker developers ship software 7x more frequently

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

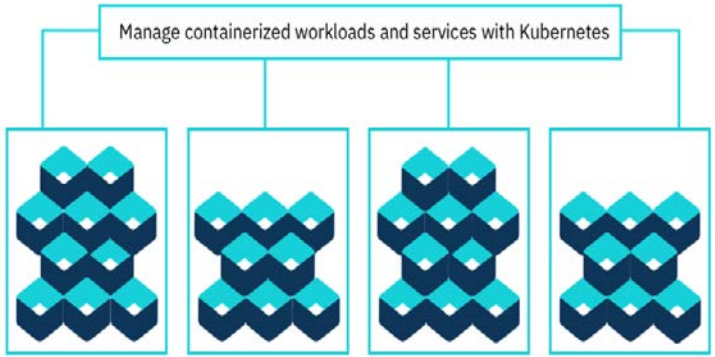
Container automation and orchestration is essential
Enter: Kubernetes

Containers are revolutionizing IT
But they require orchestration



kubernetes

Kubernetes - κυβερνήτης
Means “helmsman” or “pilot”



Private Clouds address the new IT reality

Created by digital transformation

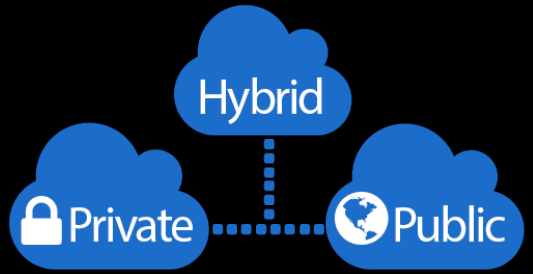
Method	Development	Deployment	Environment
 Waterfall	 Monolithic	 Bare metal	 On-Premises
 Agile Programming	 N-Tier	 Virtual Server	 Off-Premises
 Agile DevOps	 Microservices	 Containers	 Cloud

Time to value

Perception of cost

Public Cloud + Private Cloud = Hybrid Cloud *

Different cloud options



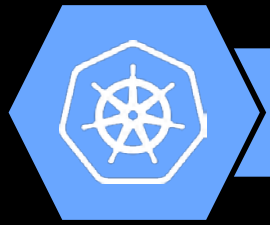
	Public Cloud	On-Premises Private Cloud	Hosted Private Cloud	Hybrid Cloud
Hardware Deployment and Management	Vendor	Customer	Vendor	Shared between vendor and customer
Hardware Sharing Model	Shared	Dedicated	Dedicated	Partially shared and partially dedicated
Scalability	High	Medium	High	High
Low Cost	Yes	Sometimes	Sometimes	Sometimes
Predictable Cost	No	Yes	Yes	No
Utility Billing	Yes	No	Depends on vendor	Partial
Flexibility	Yes	Limited	Limited	Yes
Customization Capabilities	No	Yes	Depends on vendor	Partial
Enhanced Security and Compliance	No	Yes	Yes	Yes
Instant Provisioning	Yes	Yes	Yes	Yes

* A “Hybrid Cloud” is a highly orchestrated environment, where all sources act as one

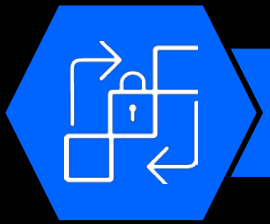
A “Multi-cloud” environment simply refers to the use of multiple cloud sources of any kind, without necessarily being orchestrated

Why care about Private Clouds?

Adoption brings agility and efficiency



Cost Efficient & Scalable Infrastructure



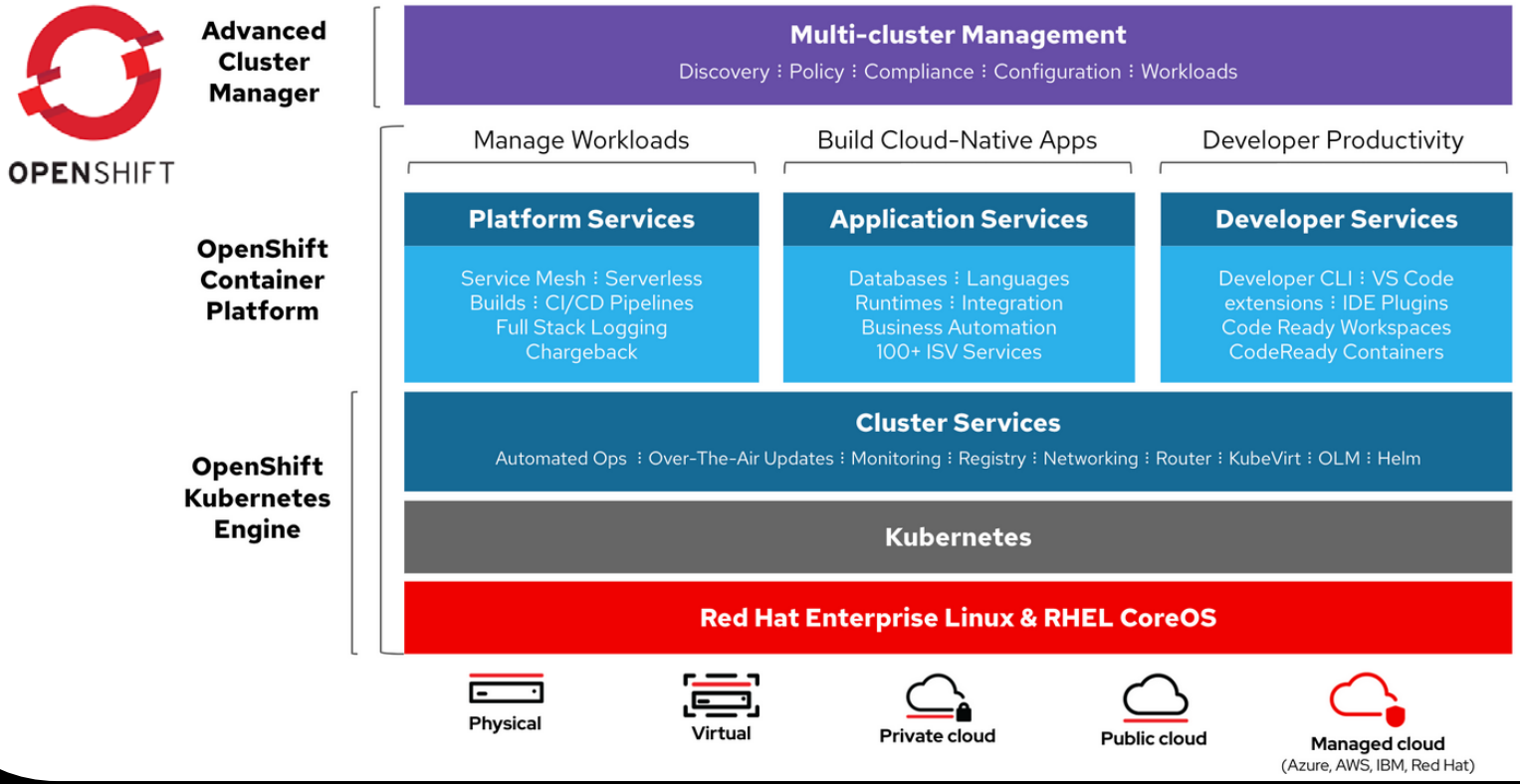
Accelerate Time to Market

Manage Data at Scale

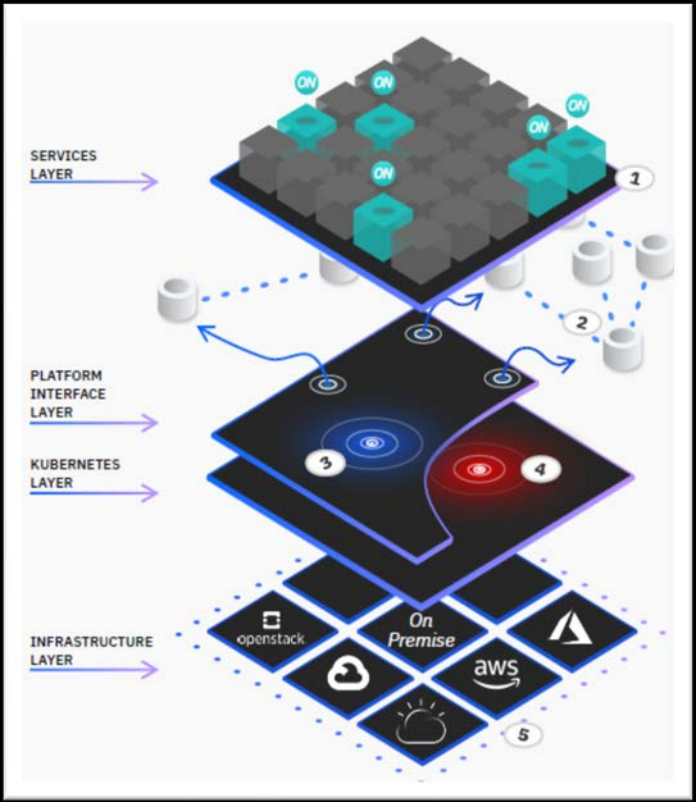
50% Benefit	**Data Center** System Utilization & Server Reduction
75% Benefit	**Manage Performance** Elasticity, Bursting, High Availability
35% Benefit	**DevOps** Faster Deployments
30% Benefit	**Deployment Efficiency** Containers & Microservices
50% Benefit	**Improved Security** Management & Risk Reduction

The Private Cloud Platform Market Leader

Red Hat OpenShift Kubernetes Engine



Cloud Pak for Data (CPD)
Architectural overview



1. Services Ecosystem

At a click, access and deploy an ecosystem of 45+ services and templates from IBM and third parties.

2. Data Virtualization

Query across multiple data sources fast and easy without moving your data.

3. Control Plane (Platform Interface)

Speed time to value with a single platform that integrates data management, data governance and analysis for greater efficiency and improved use of resources

4. OpenShift (Kubernetes Engine) RED HAT OPENSHIFT

Leverage the leading hybrid cloud, enterprise container platform for an innovative and fast deployment strategy.

5. Any Cloud (Infrastructure)

Avoid lock-in and leverage all cloud infrastructure with our “Any Cloud” mentality.

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)

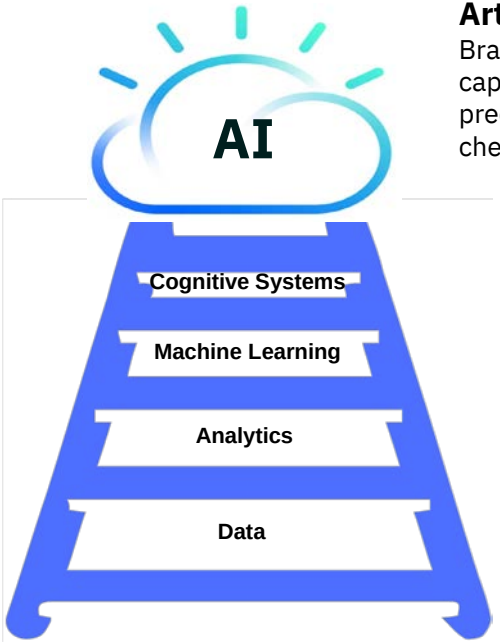
2. Premium Services *(purchase license or BYOL)*

Collect	Organize	Analyze	Deploy / Infuse
Db2 AESE Virtual Data Pipeline	Infosphere DataStage Edition Infosphere Regulatory Accelerator Infosphere multi-cloud Data Mvmnt 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

mongoDB. · Datameer · accelerite · Prolifics · figure eight · Lightbend · Cockroach LAB · NetApp portworx · EDB POSTGRES · wand · OA4ERP · PlanetScale · DR · Senzing · knowis · ANALYTICS ZOO · OA4ERP

The Original* AI Ladder
Described: AI vs. Cognitive systems vs. Machine Learning vs. Analytics



The “AI Ladder”

Artificial Intelligence
Branch of computer science dealing with the simulation of intelligent behavior in machines. It is the capability of machines imitating intelligent (usually human) behavior. AI has been around for many years, predating machine learning and the term “cognitive systems.” e.g. IBM Deep Blue beat the world champ at chess in 1996 by using brute force computer programming.

Cognitive systems
Simulation of human thought processes in a computerized model with self-learning systems that use data mining, pattern recognition and natural language processing to mimic the way the human brain works. e.g. IBM Watson beats Jeopardy! Champs in 2011 using an ML based predictive analytics driven by NLP. Modern AI applications can also be described as “Cognitive.”

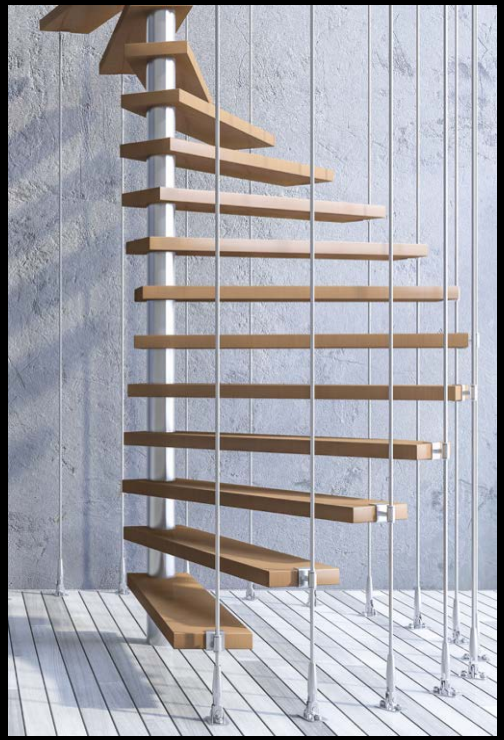
Machine Learning
Branch of AI that employs statistical models and advanced algorithms to enable computers to become "intelligent" by "learning" from data in lieu of being specifically programmed. Can include predictive analytics, NLP, facial recognition, search engines. e.g. Google’s DeepMind beats Go champion by using ML neural networks and search.

Analytics
The discovery, interpretation, and communication of meaningful patterns in data.

Data
Includes structured, semi-structured and unstructured datatypes in on-premises, public or private cloud, and hybrid environments.

* Presented in 2017 by Rob Thomas, IBM Senior Vice President, Cloud and Data Platform

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

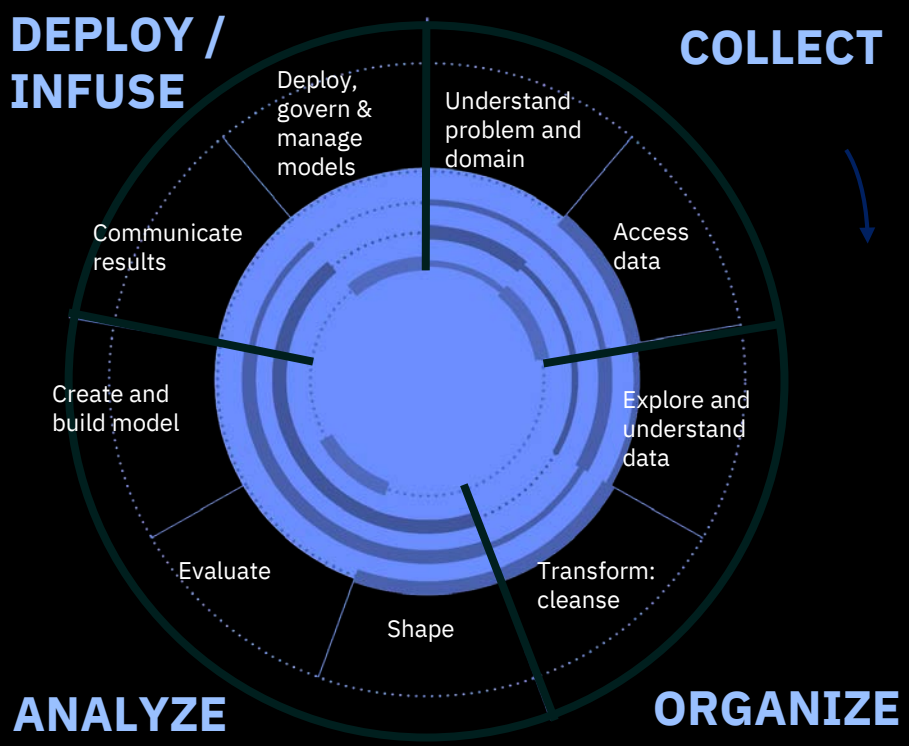
Organize – Create a trusted analytics foundation

Collect – Make data simple & accessible

Strong Foundation – Built on "Cloud native architecture"

Cloud Pak for Data (CPD)

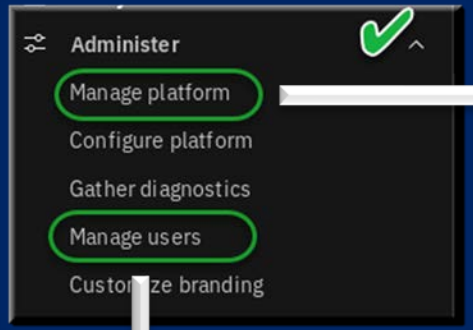
Increases workforce productivity across the analytics lifecycle



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



Deployments

See the deployments running in each Red Hat OpenShift project on the platform. You can see how many resources each deployment is currently using compared to the number of resources that the deployment has reserved.

Filter by: All types | Clear all

Find deployments

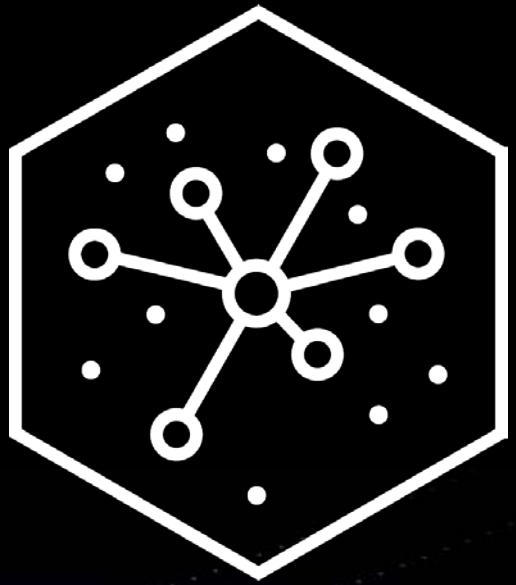
Name	Type	Installed on	Service instances	vCPU	Memory (GB)
Total			2	6.01 of 75.08	101.20 of 201.70
IBM Cloud Pak for Data Cognos Analytics Advanced	Premium	5/28/2020 4:32 PM	0	0.00 of 0.60	0.11 of 0.11
SPSS Modeler for Watson Studio	Premium	5/28/2020 4:32 PM	0	1.45 of 2.10	0.42 of 8.40
Db2 Warehouse	Base	5/28/2020 4:32 PM	1	0.67 of 10.15	16.80 of 44.10

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.

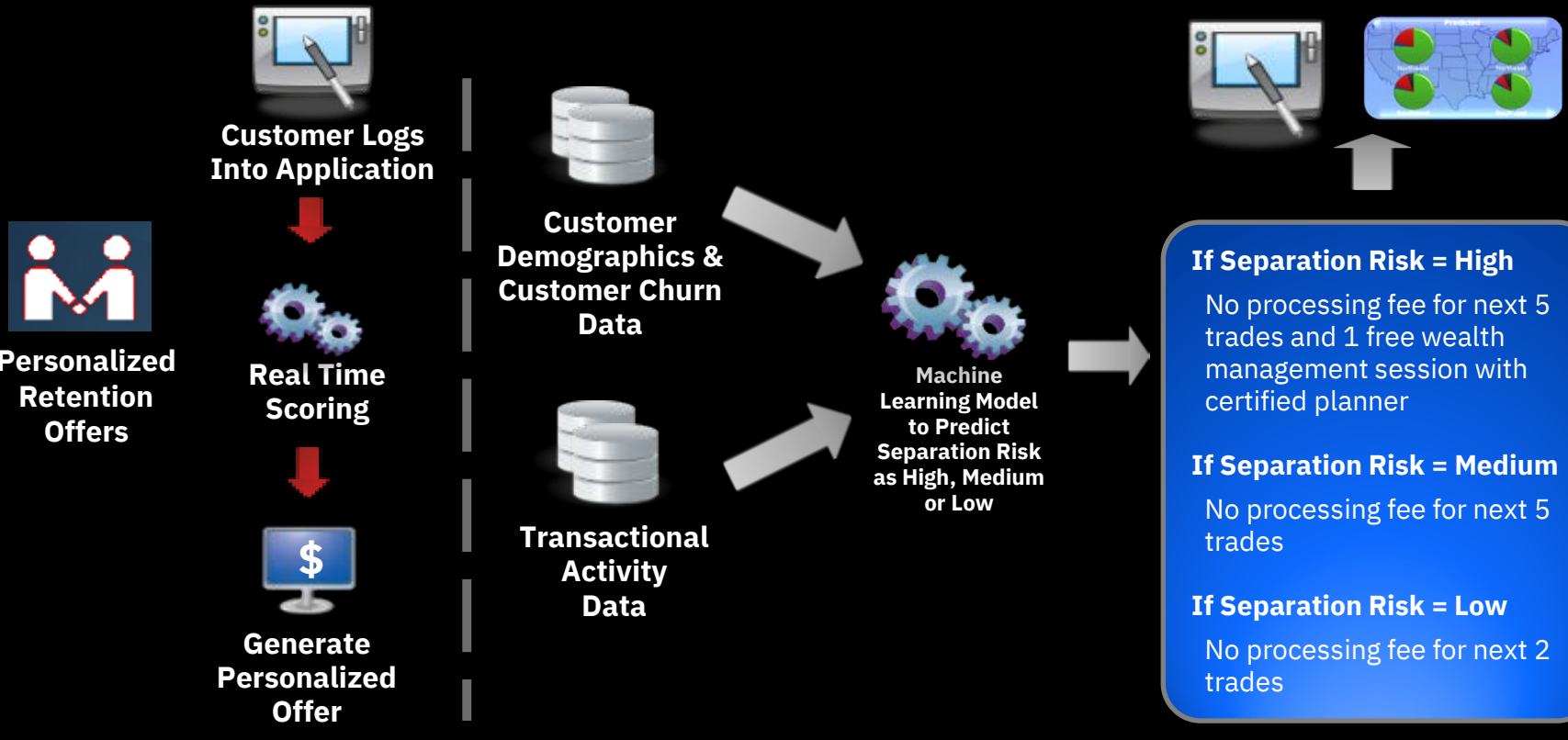
Data Driven Process to Develop Machine Learning Models

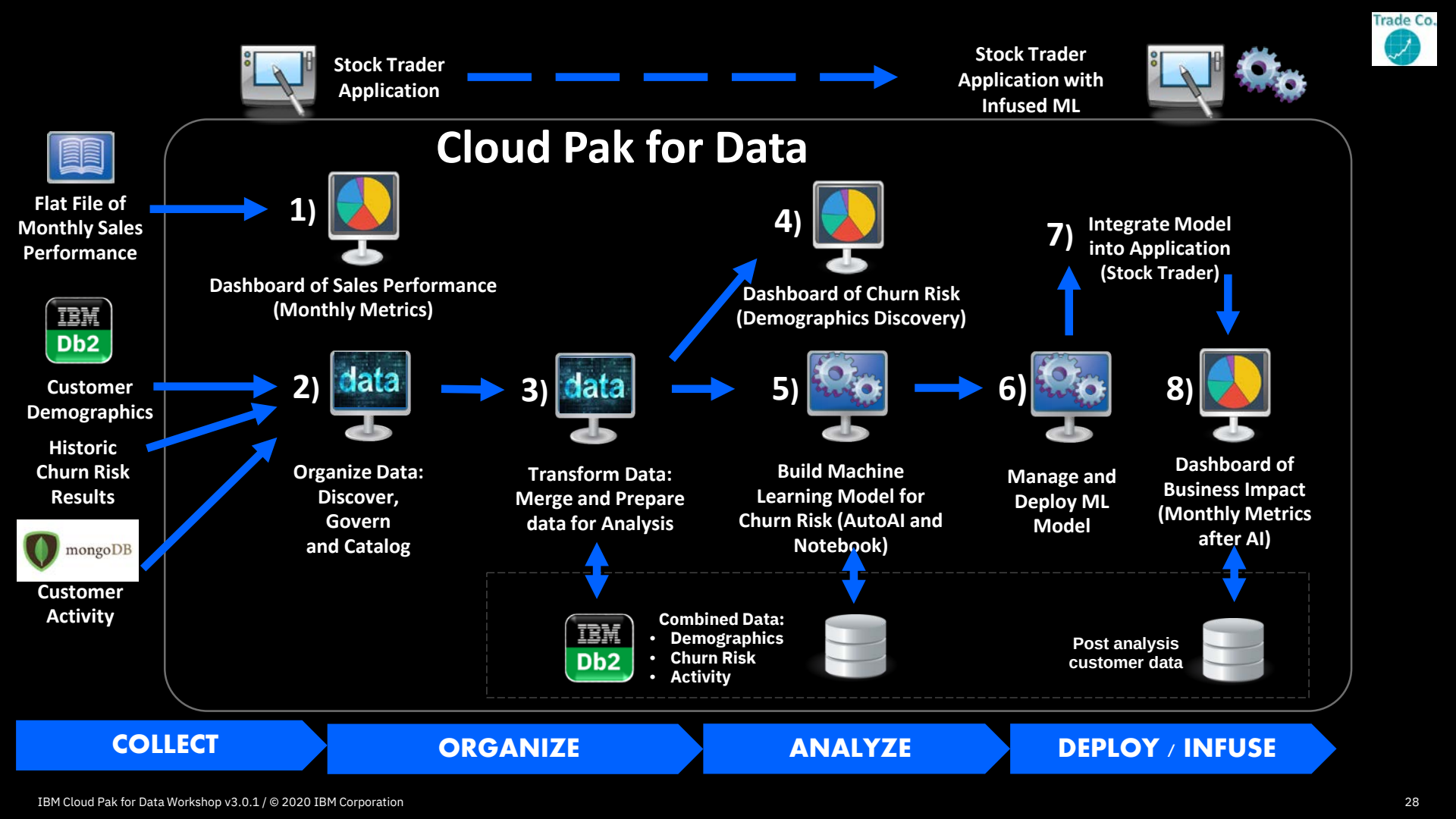
Develop predictive models for separation risk that automatically discover and incorporate all the patterns in the data including interactions and contingencies.

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.

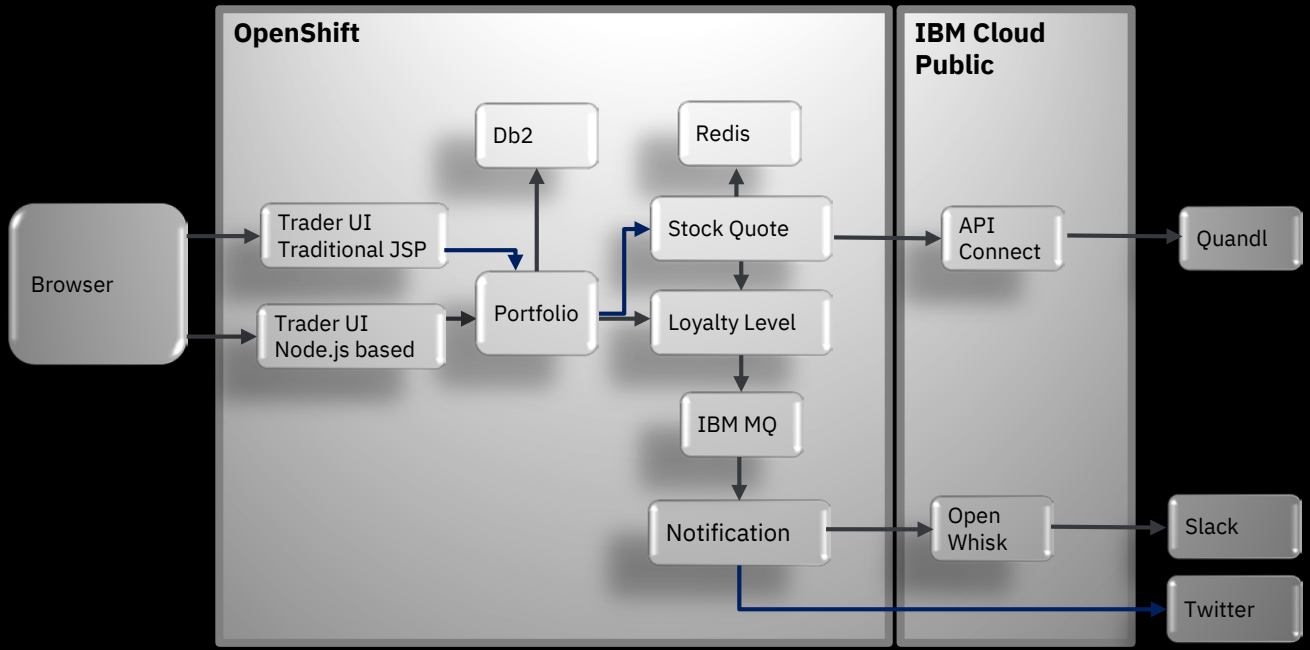
Deployment: Stock Trader App. Integrated with AI




Stock Trader Application
Stock Trader Application with Infused ML
Cloud Pak for Data
Flat File of Monthly Sales Performance
1)
Dashboard of Sales Performance (Monthly Metrics)
4)
Dashboard of Churn Risk (Demographics Discovery)
7)
Integrate Model into Application (Stock Trader)
Customer Demographics
Historic Churn Risk Results
Customer Activity
2)
Organize Data: Discover, Govern and Catalog
3)
Transform Data: Merge and Prepare data for Analysis
5)
Build Machine Learning Model for Churn Risk (AutoAI and Notebook)
6)
Manage and Deploy ML Model
8)
Dashboard of Business Impact (Monthly Metrics after AI)
Combined Data:
Demographics
Churn Risk
Activity
Post analysis customer data
COLLECT
ORGANIZE
ANALYZE
DEPLOY / INFUSE

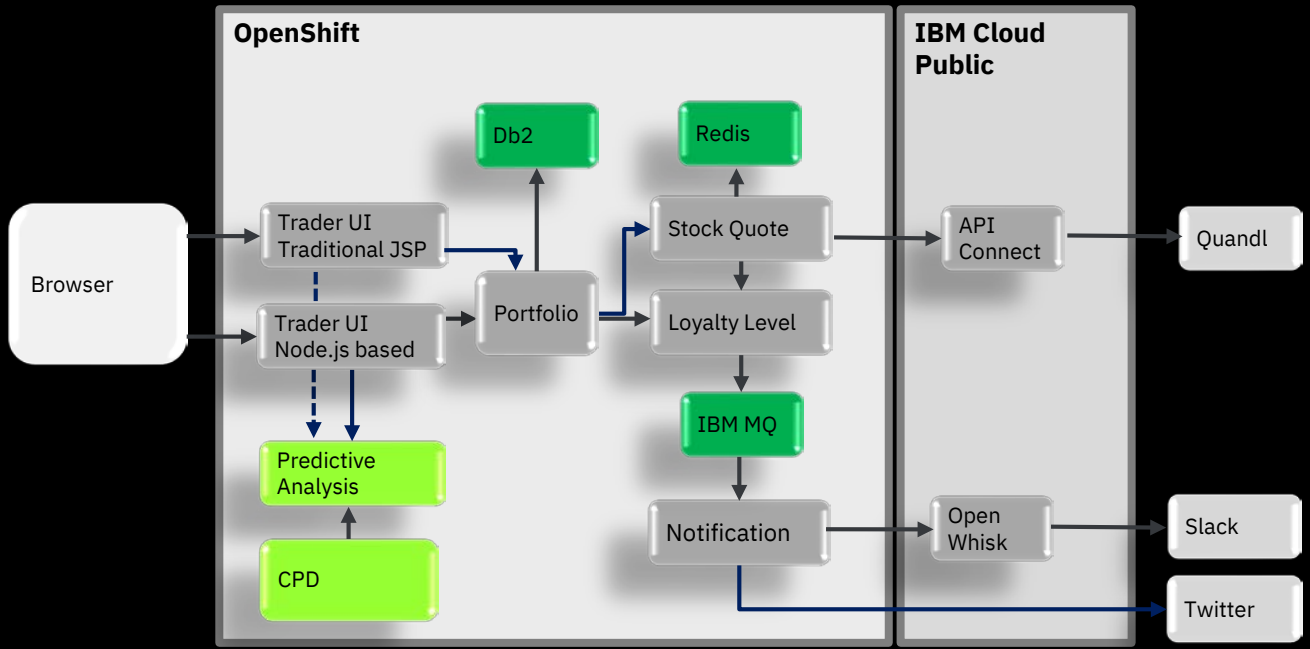
Trade Co. Application “Stock Trader” – Before
Microservices Application

Stock Trader
as a modern microservice application

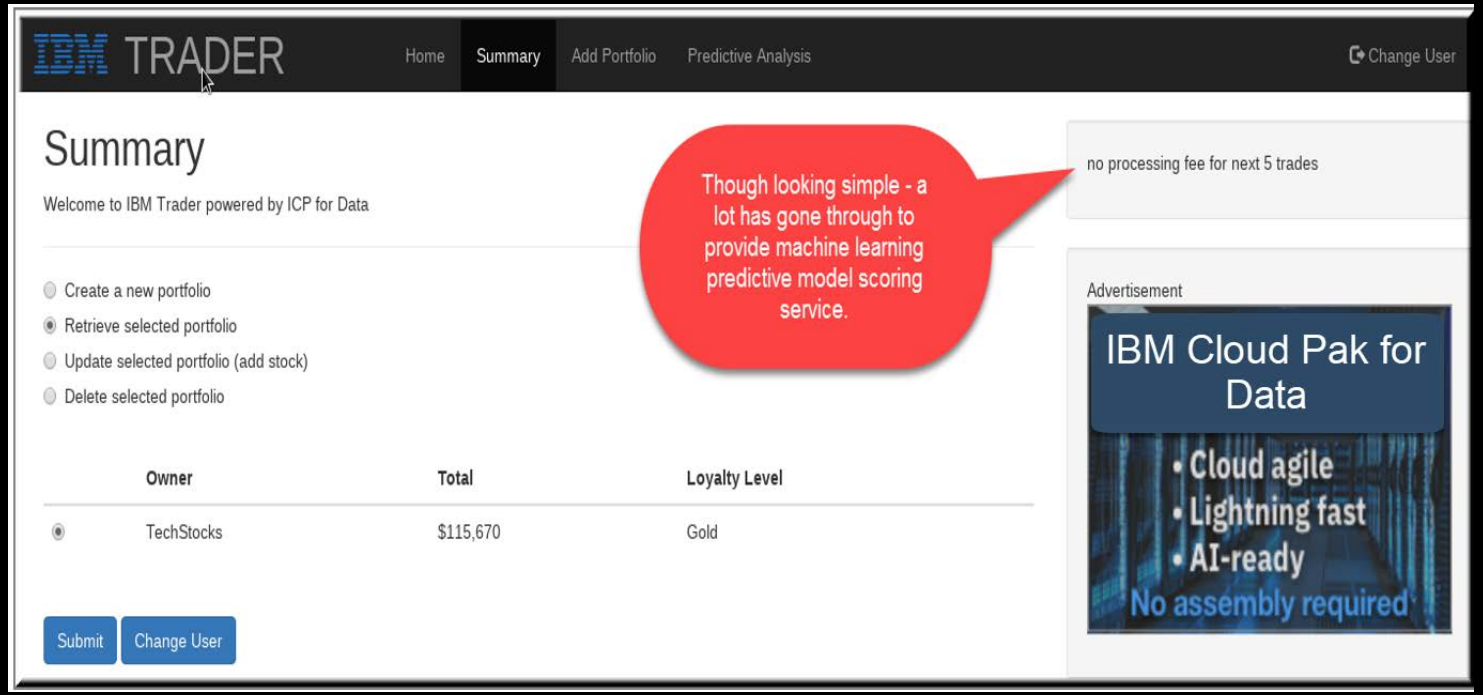


Trade Co. Application “Stock Trader” – After: Infused with ML
Microservices Application

Stock Trader (enhanced)
as a modern microservice application



Stock Trader – After
Monetizing the ML model



Trade Co. Dashboards

Before and After deploying the CPD developed ML model

Before AI

After AI

IBM Analytics Modernization Workshop

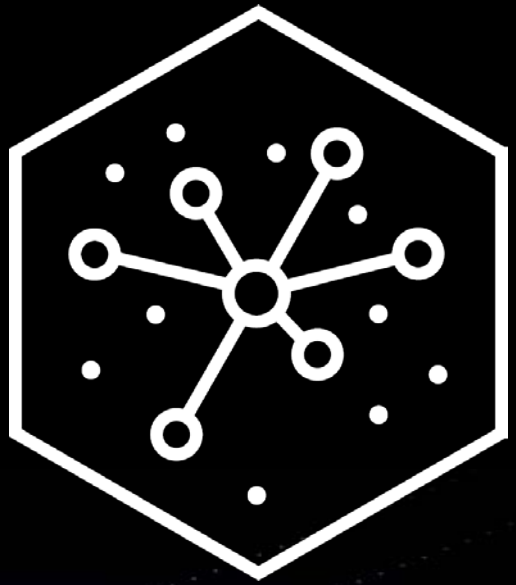
Part 1

• Introduction • Business Use Case	• Lab 01 • Lab 02
• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

IBM Analytics Modernization Workshop

Part 2

• Introduction • Business Use Case	• Lab 01 • Lab 02
• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

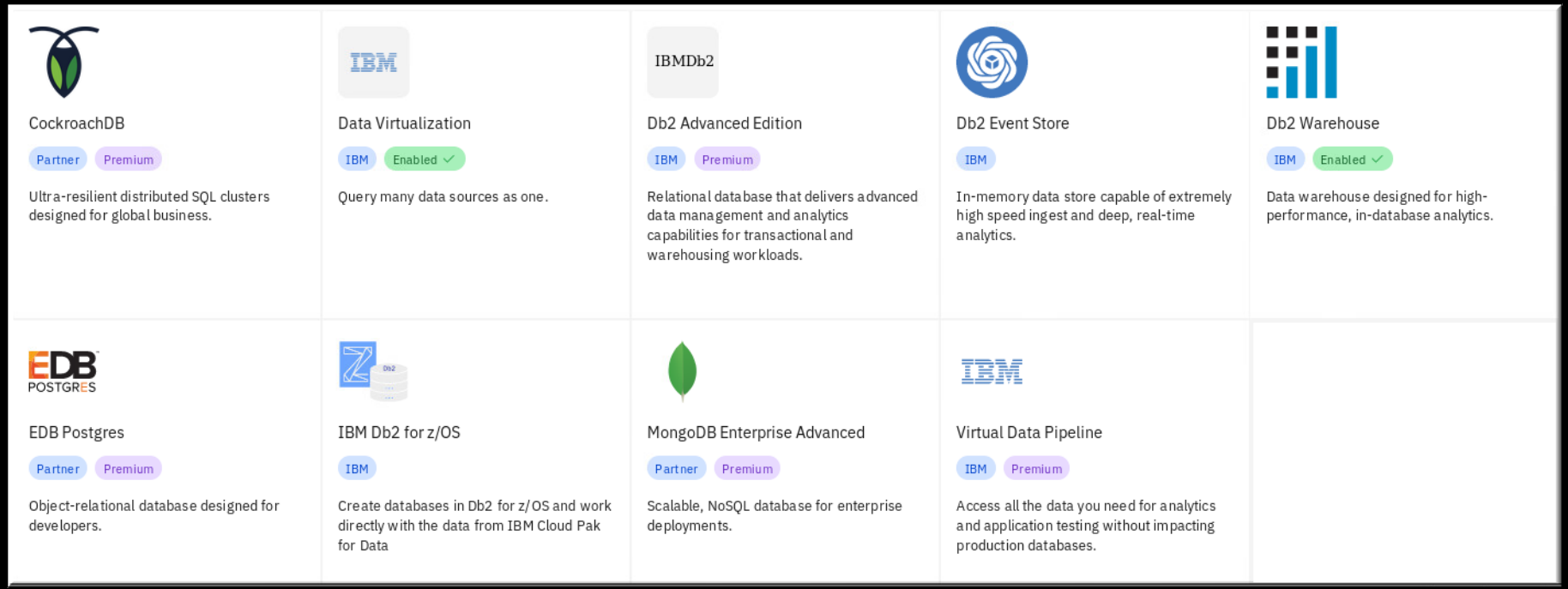


Collect

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



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
Amazon Redshift			✔	✔	✔
Amazon S3		✔		✔	✔
Apache Cassandra		✔			
Apache Derby			✔		
Apache Hbase		✔			
Apache HDFS				✔	✔
Apache Hive			✔	✔	✔
Apache Kafka		✔			
Azure Storage		✔			
BDFS		✔			
Cloudera Impala			✔	✔	✔
Dropbox				✔	✔
Filesystem		✔		✔	✔
FTP Enterprise		✔			
FTP		✔		✔	✔
Generic JDBC				✔	✔
Google BigQuery		✔	✔	✔	✔
Google Cloud Storage		✔		✔	✔
HDFS Generic web-HDFS				✔	
HDFS HttpFS				✔	
Hive JDBC		✔	✔		
Hive JDBC CDH		✔		✔	
Hive JDBC HDP		✔		✔	
Hortonworks HDFS				✔	✔

Third-party Data Sources	Cognos Dashboards	DataStage Edition	Data Virtualization	WKC	Watson Studio
Looker				✔	✔
MariaDB			✔		
MSFT Azure Blob & File		✔			
MSFT Azure Lake Store				✔	✔
MSFT Azure SQL DB				✔	✔
MSFT SQL Server	✔	✔	✔	✔	✔
Minio				✔	✔
Mongo			✔	✔	
MySQL			✔	✔	✔
ODBC		✔			
OData				✔	✔
Oracle		✔	✔	✔	✔
Pivotal Greenplum		✔		✔	✔
PostgreSQL	✔		✔	✔	✔
Salesforce.com		✔		✔	✔
SAP HANA			✔		
SAP Data Object		✔		✔	✔
Snowflake		✔	✔	✔	✔
Sybase Enterprise		✔	✔	✔	✔
Sybase IQ / OC		✔		✔	✔
Tableau				✔	✔
Teradata		✔	✔	✔	✔

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:

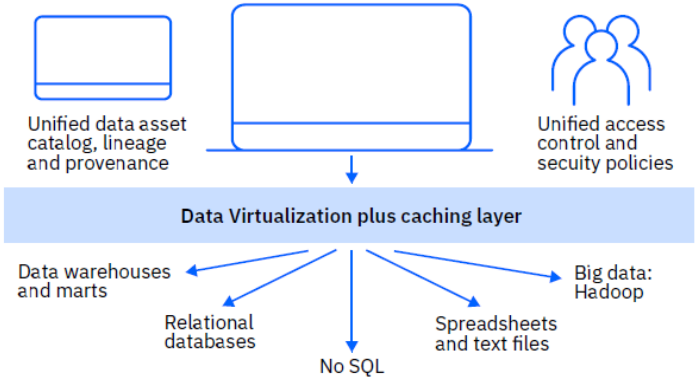
Service	Comment
Cognos Dashboards	You can use the local and remote data sets that already exist in your analytics projects. You can create connections by selecting *Add data* source from the analytics dashboard menu.
DataStage Edition	You can transform data that is in a catalog by searching for the data that you want to use and selecting *Transform*.
Data Virtualization	You can create connections that can be used to virtualize data from the following locations: 1) *Connections* page, 2) *Data Sources* 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) *Connections* page, 2) *Assets* page of the analytics project. You can also *Add data* from files.

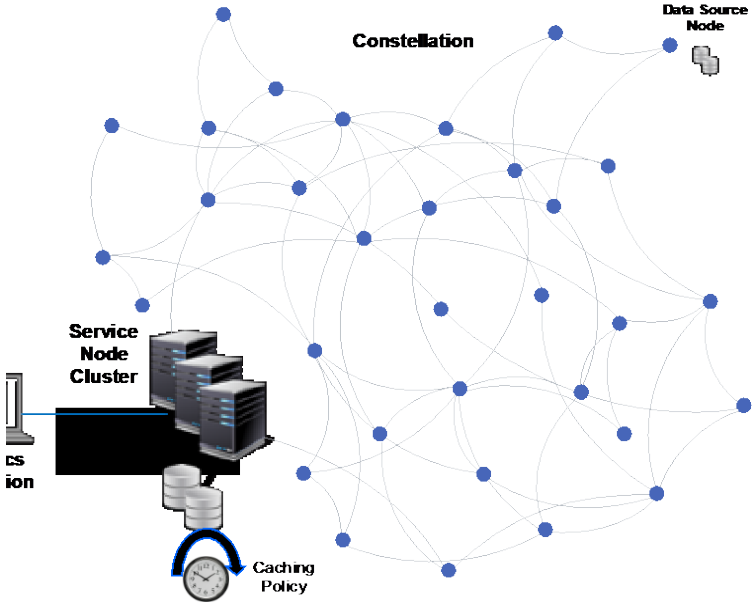
CPD Collect
3. Data Virtualization - Overview

Query across multiple databases and big data repositories which appear as one to an application

Data Virtualization in Cloud Pak for Data is a unique new technology that connects many data sources into a single self-balancing collection of data sources or databases referred to as a *constellation*

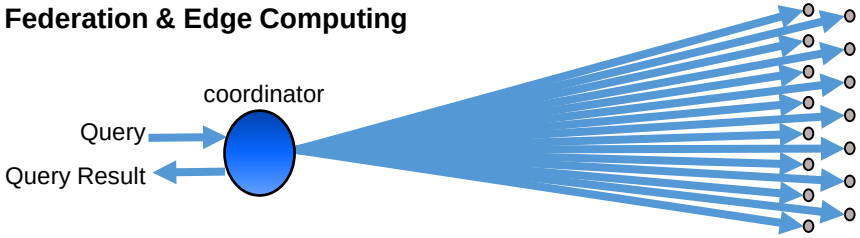
- Analytics are submitted against the data at the source (data is not copied)
- New algorithms are applied that support distributed SQL execution across heterogenous data sources





CPD Data Virtualization
Constellation “Computational Mesh” benefit

Classic Federation & Edge Computing

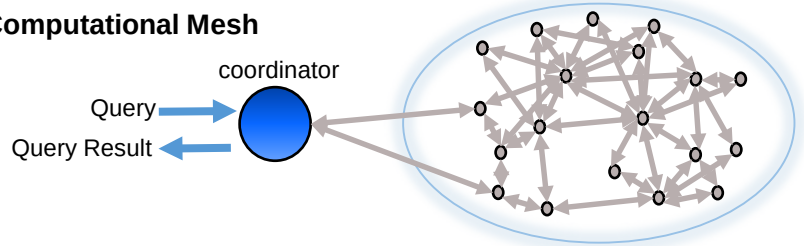


Query issued against the system	A coordinator receives the request and fans the work out to edge nodes	Edge nodes individually perform as much work as they can based on their own data. Individual results are sent back to the coordinator for final merging and remaining analytics.	Coordinator receives 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

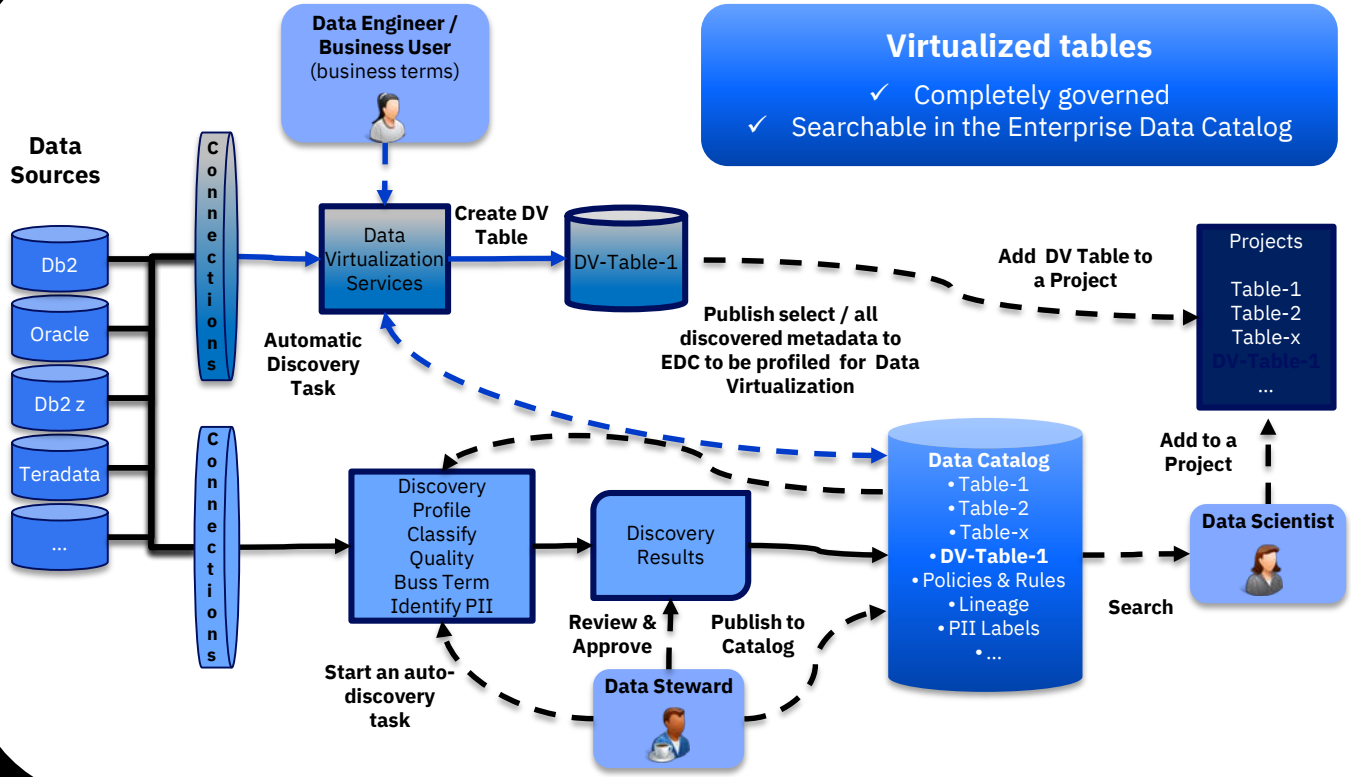
New Computational Mesh



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.

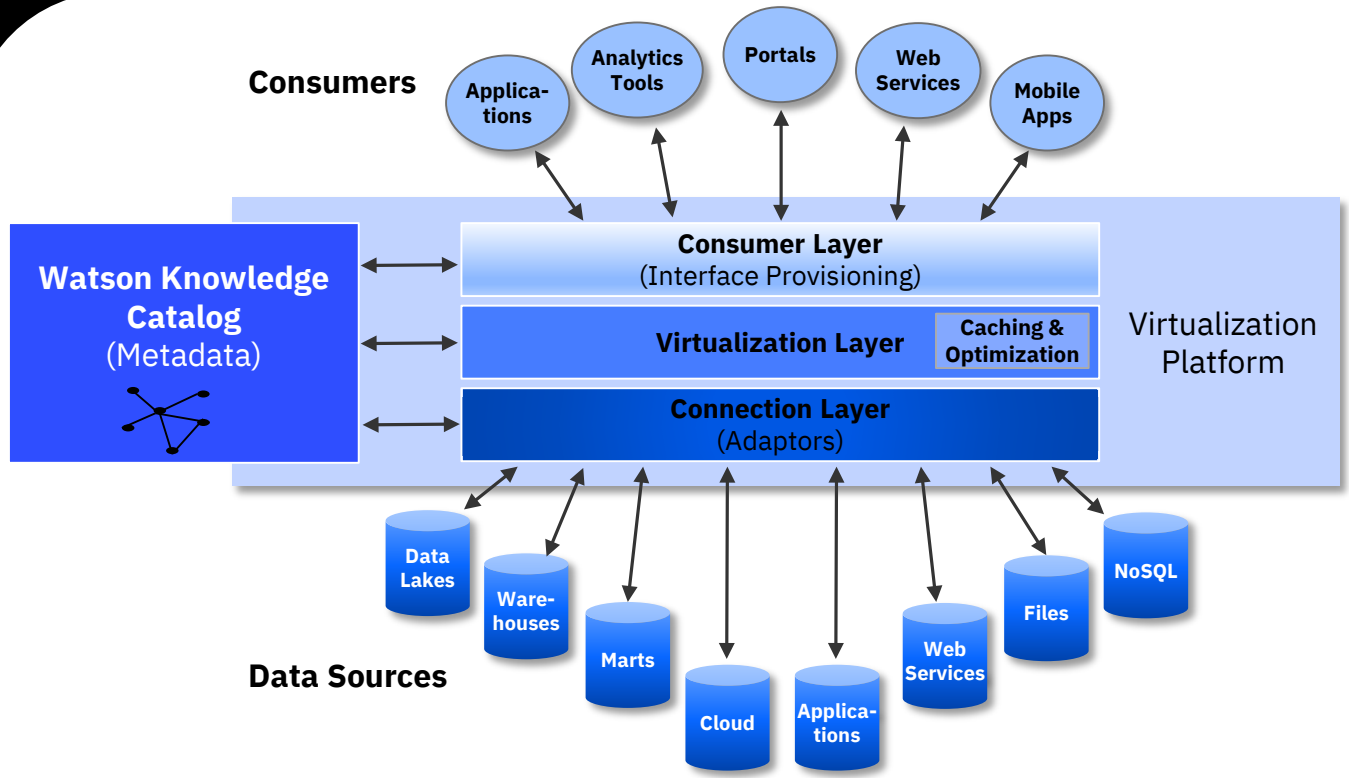
CPD Data Virtualization
Data Governance is built in



- By using this flow, the DV Table will be published to catalog with fingerprint information collected (quality, profile, classification, assign business terms, Sensitive Information Detection)
- If any PII data is present, the system will automatically flag the table in the catalog with a PII Label
- Data Masking Services can de-identify the sensitive information when user attempts to view the data

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

CPD Collect
4. External Data Sets

The Weather Company

Historical Weather Data

- 3 Years of Historical Weather Data
- Current Weather Conditions
- Weather Forecast Data
- Location Look-up Services
- Industry Accelerators (Retail & Manufacturing)

90-day trial

Equifax

Demographic Consumer Data

- Ability to Pay
- Economic Cohorts
- Credit Styles Pro
- Income 360
- Wealth Complete Data
- IXI-Data

Premium

People Data Labs

People Data

- Dataset of ~1.5B profiles consisting of both b2b and b2c data on each person
- Data is accessible via an Enrichment API or Data License

Premium

BCC

Real Time Stock Data

- Real Time Stock Market Data

Premium

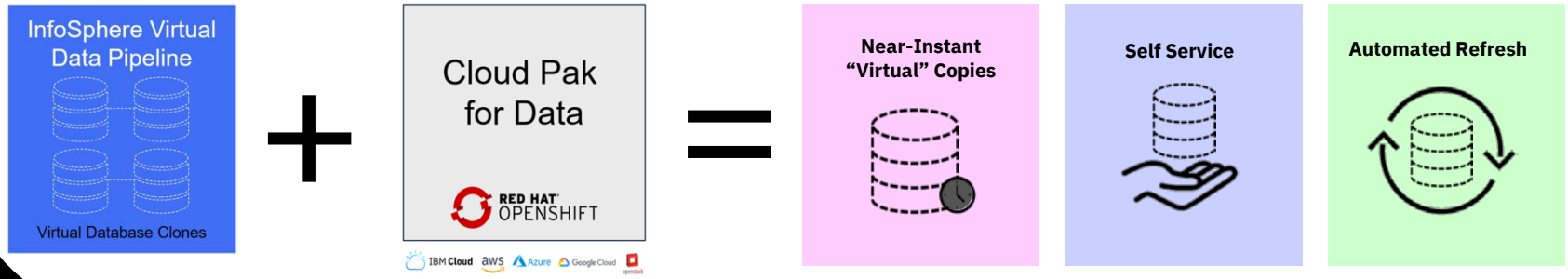
CPD Collect
5. Premium Service: Virtual Data Pipeline – Overview

Provision and refresh analytics and test data in minutes

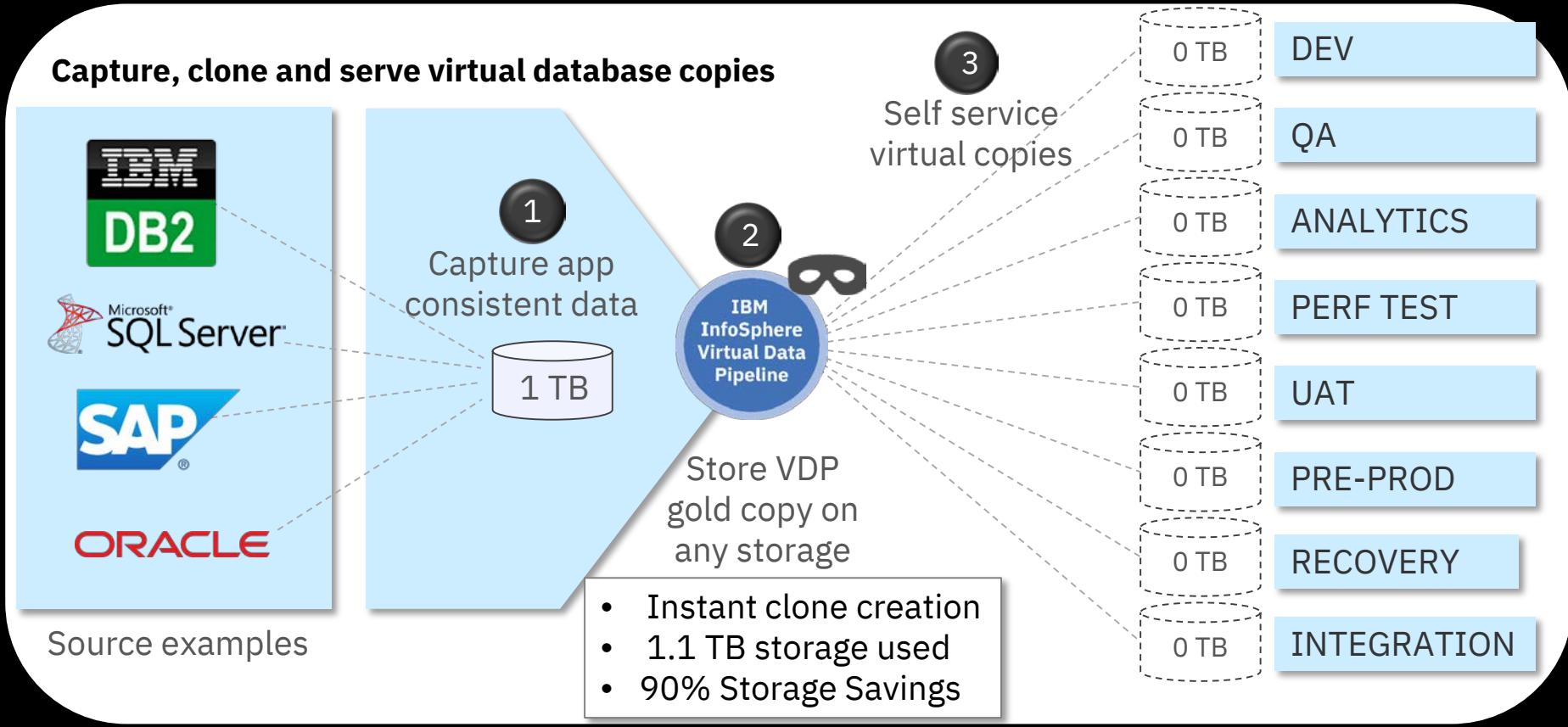
Virtual Data Pipeline in Cloud Pak for Data is a service that allows users to *instantly provision virtual database copies* to work with near real-time for data analytics and application testing, AI model training and testing, and data virtualization.

Accessing production data quickly and securely can be a challenge:

- *Risk*: Accessing Production data can impact operations and present a security and data privacy risk
- *Time Consuming*: Moving data to data warehouses and marts or creating duplicate copies can result in stale data
- *Cost*: Creating multiple copies and versions of data for each functional area uses a lot of storage and increases costs

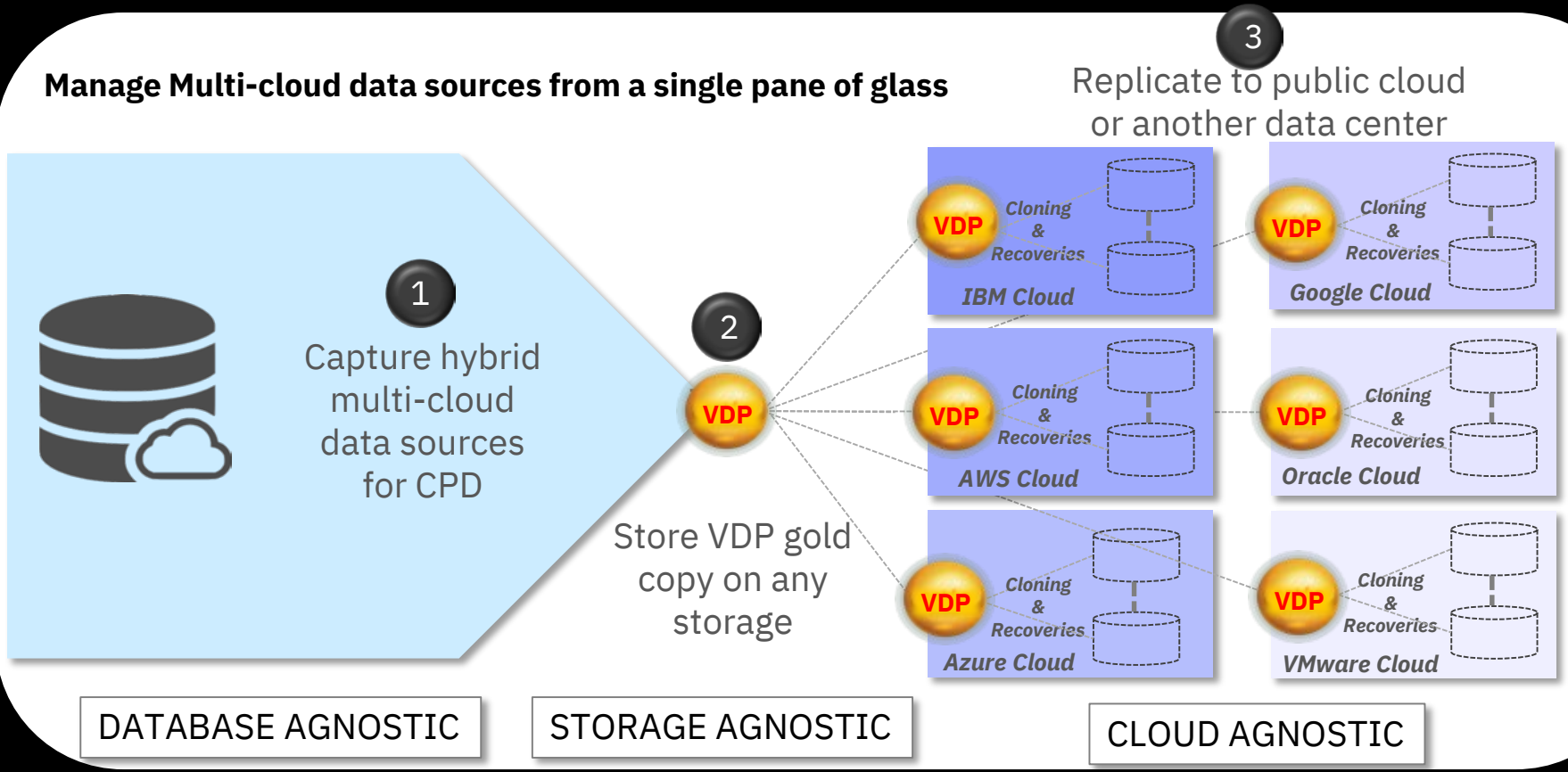


Virtual Data Pipeline
Cost savings example



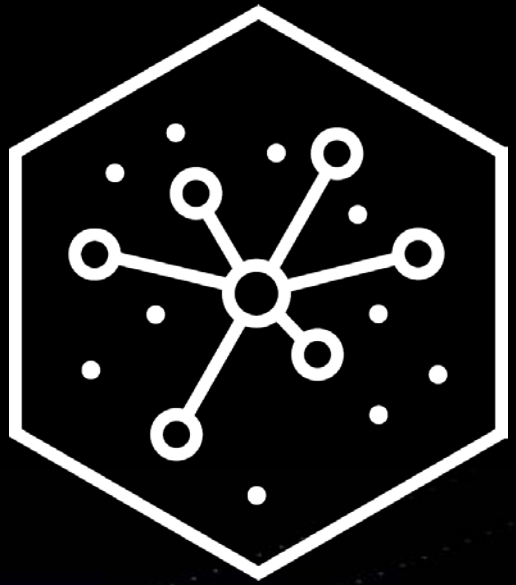
Virtual Data Pipeline
Multi-cloud data management example

Manage Multi-cloud data sources from a single pane of glass



Organize

Lab 04 – Organize



CPD Organize
Watson Knowledge Catalog (WKC)

Data Scientists
Data Analysts
Business Analysts

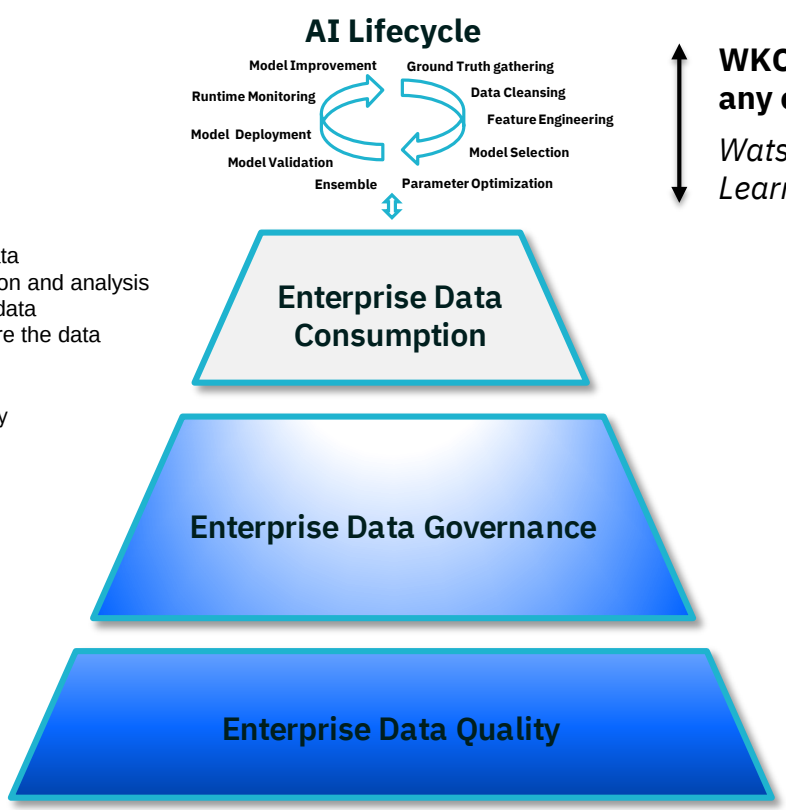
- Search and find relevant data
- Prepare data for consumption and analysis
- Consume and analyze the data
- Rate, comment on and share the data

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



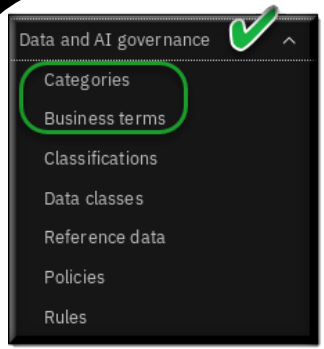
WKC is installed when you install any of the following services:

Watson Studio, Watson Machine Learning, Watson OpenScale

IBM Watson Knowledge Catalog on Cloud Pak for Data

End-to-End Platform for Business-Ready Data

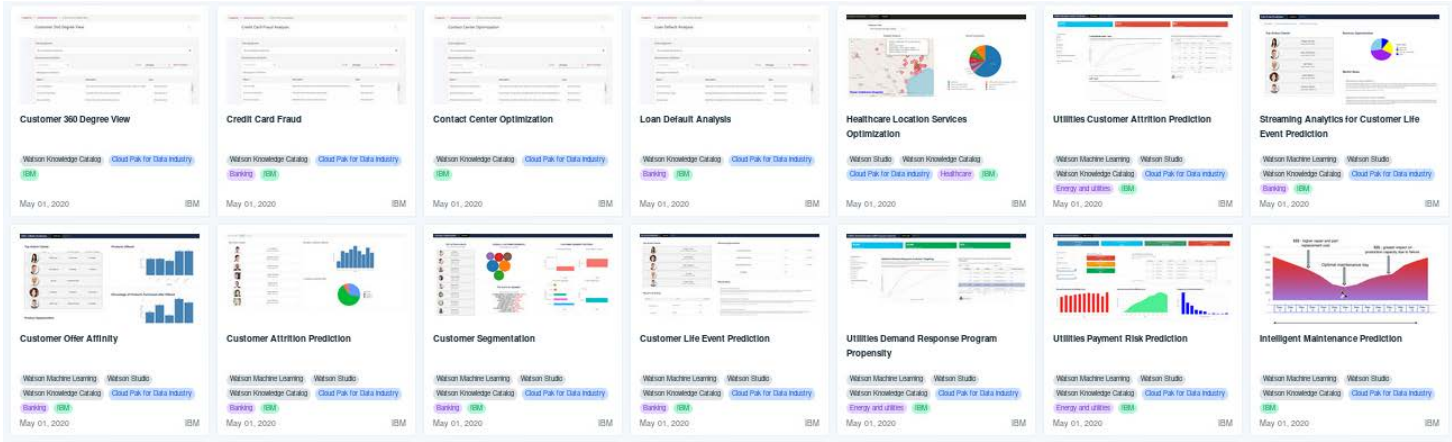
CPD Organize
Data and AI governance: Categories and Business terms



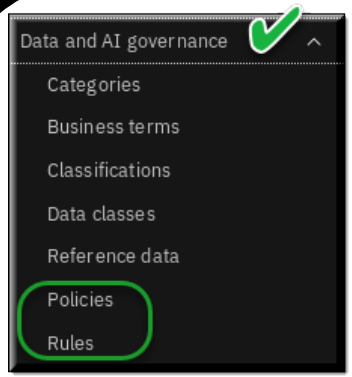
Categories provide logical structure to a business glossary

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**

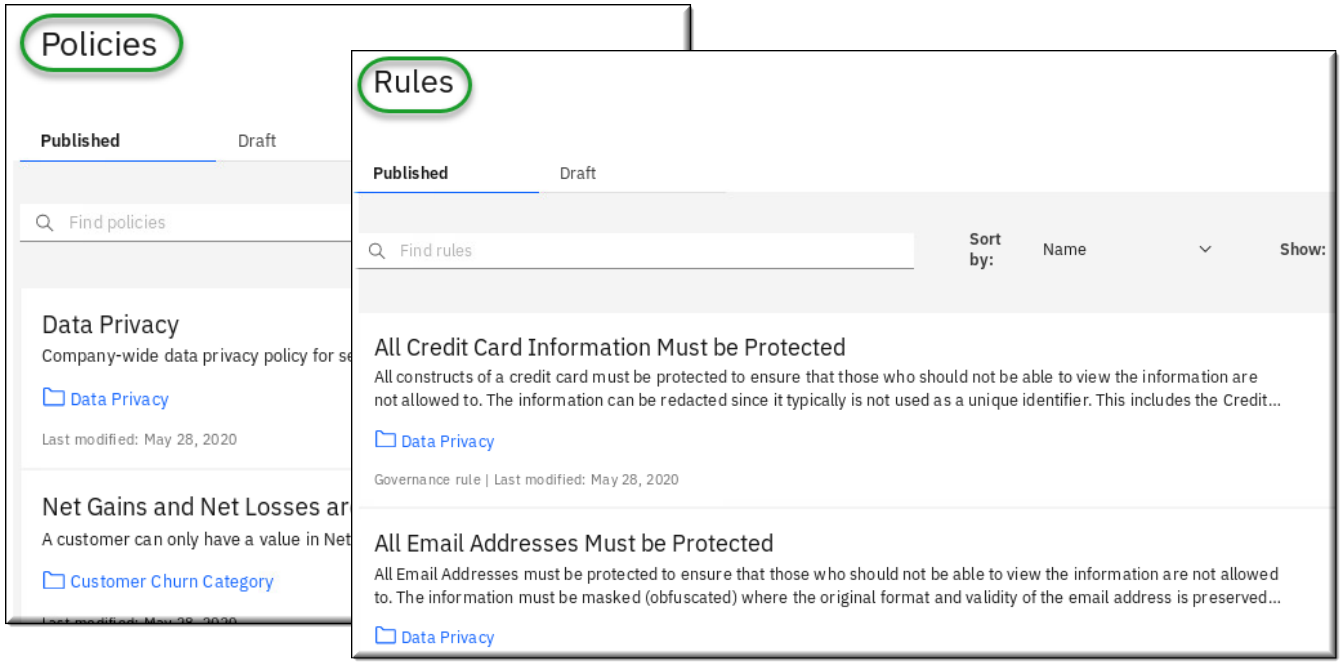


CPD Organize
Data and AI governance: Policies and Rules

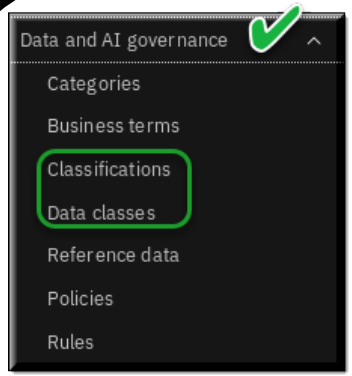


Policies describe how to control data and consist of one or more rules

Rules describe the criteria for compliance with business objectives

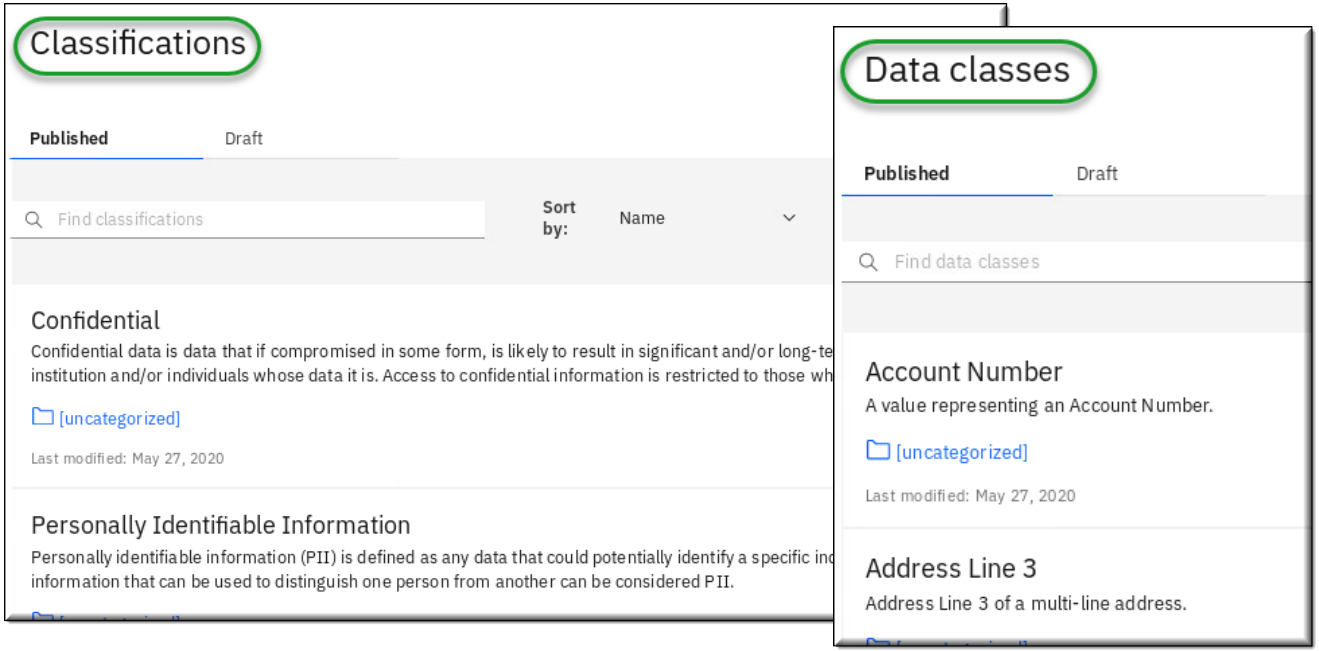


CPD Organize
Data and AI governance: Classifications and Data classes

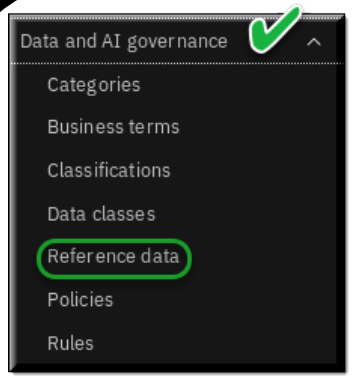


Classifications describes the sensitivity level of data

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

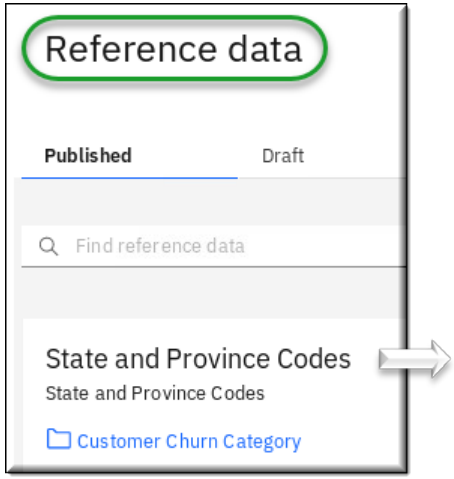


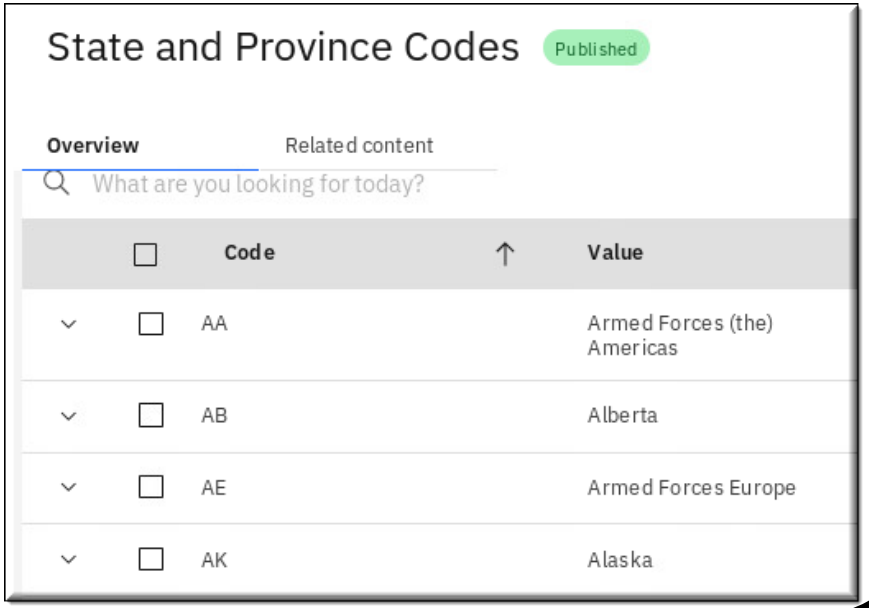
CPD Organize
Data and AI governance: Reference Data



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

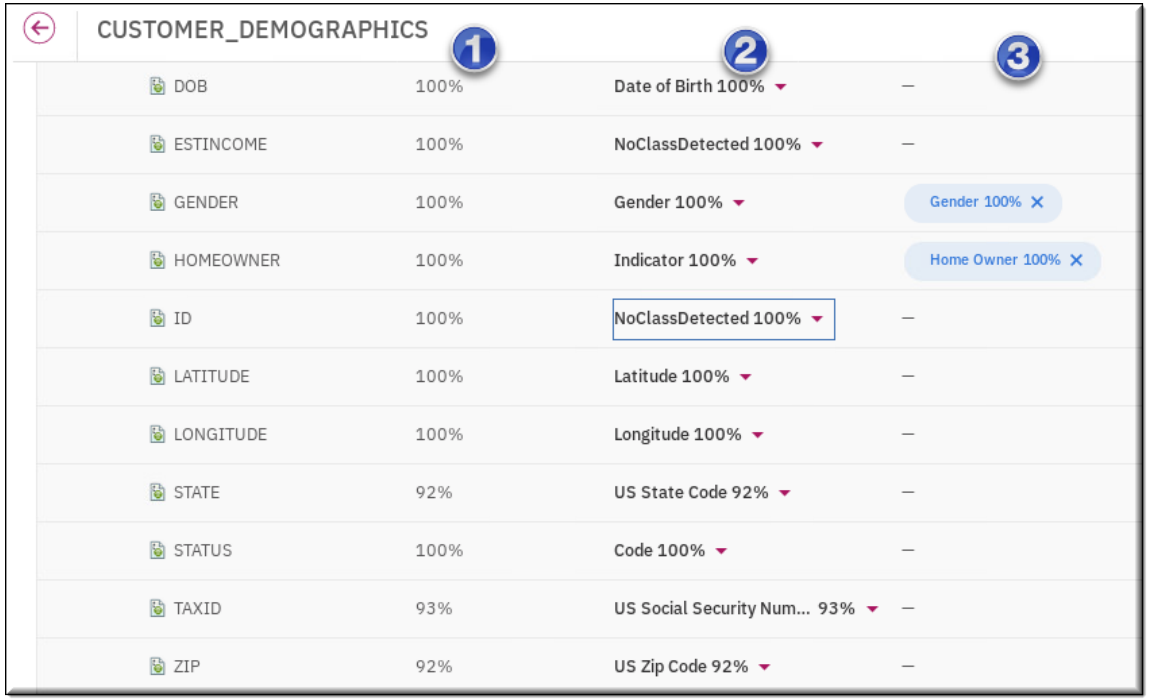
May be referenced by Business Terms, Policies, Rules and Data Classes





CPD Organize
Auto-discover assets

Data Discovery



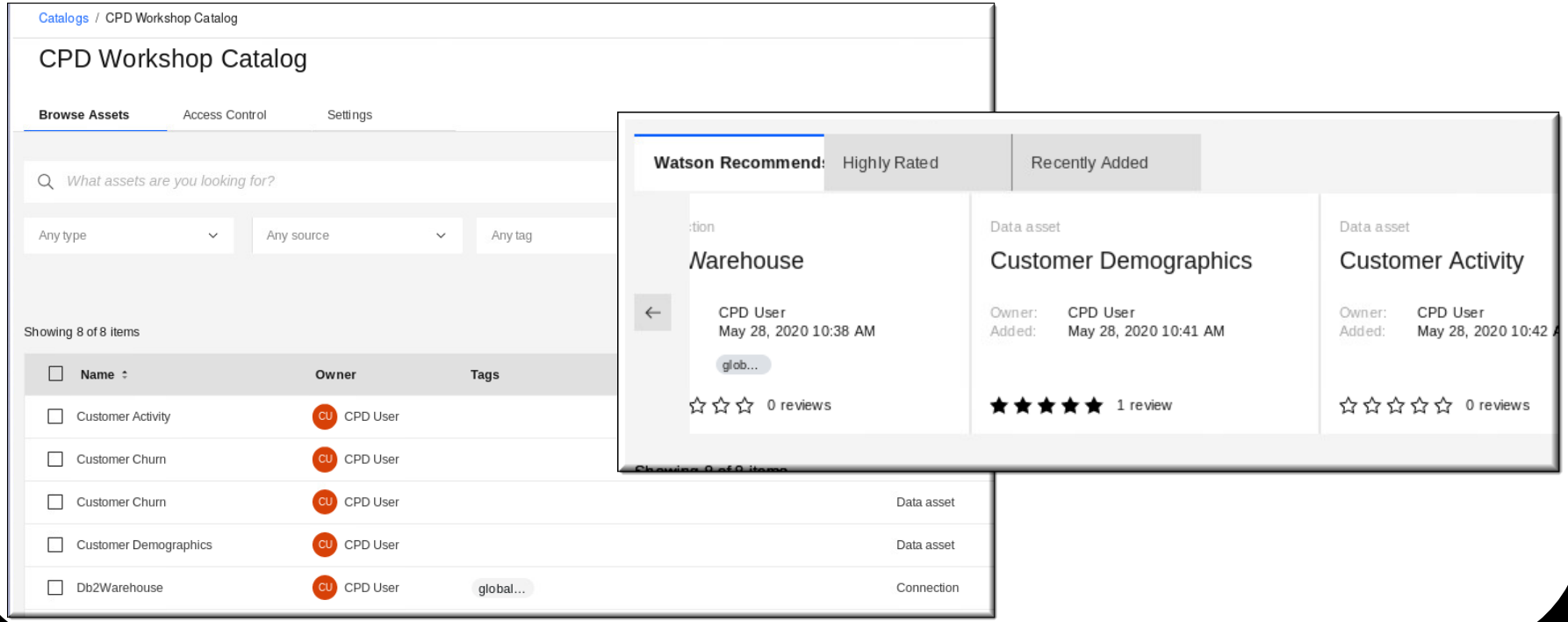
Use machine learning based auto-discovery to:

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.

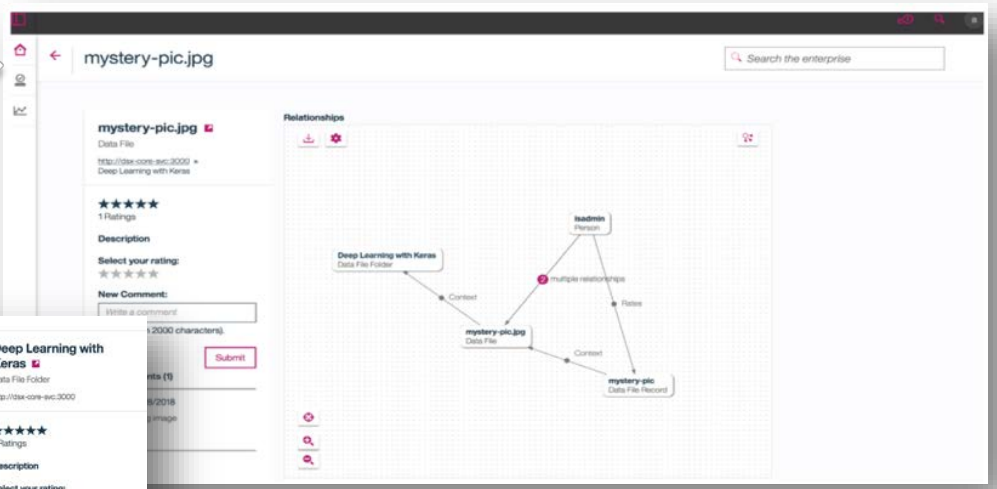
CPD Organize
Publish to a catalog

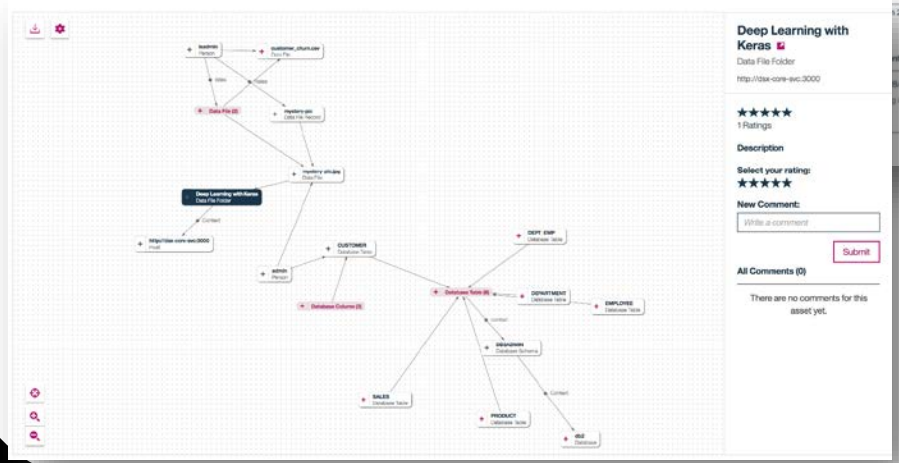
Catalog and govern your assets



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





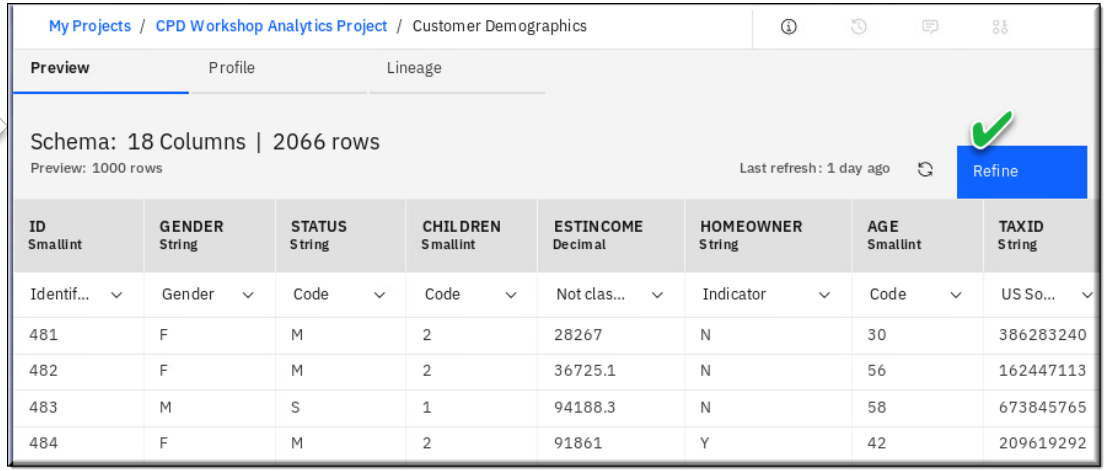
Explore deeper to understand context and usage patterns

CPD Organize
Refine data with visualizations

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

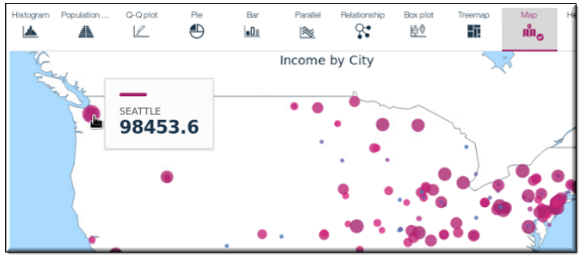
Use it to remove data that is incorrect, incomplete, improperly formatted, etc.

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.

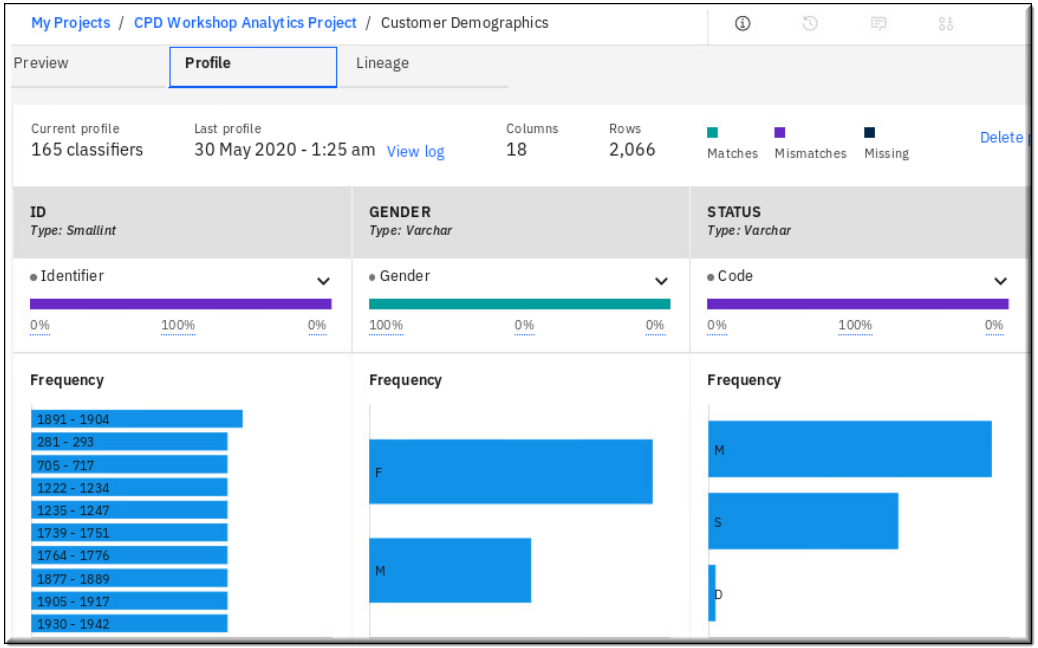
All catalog or project members can see data asset profiles.

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:

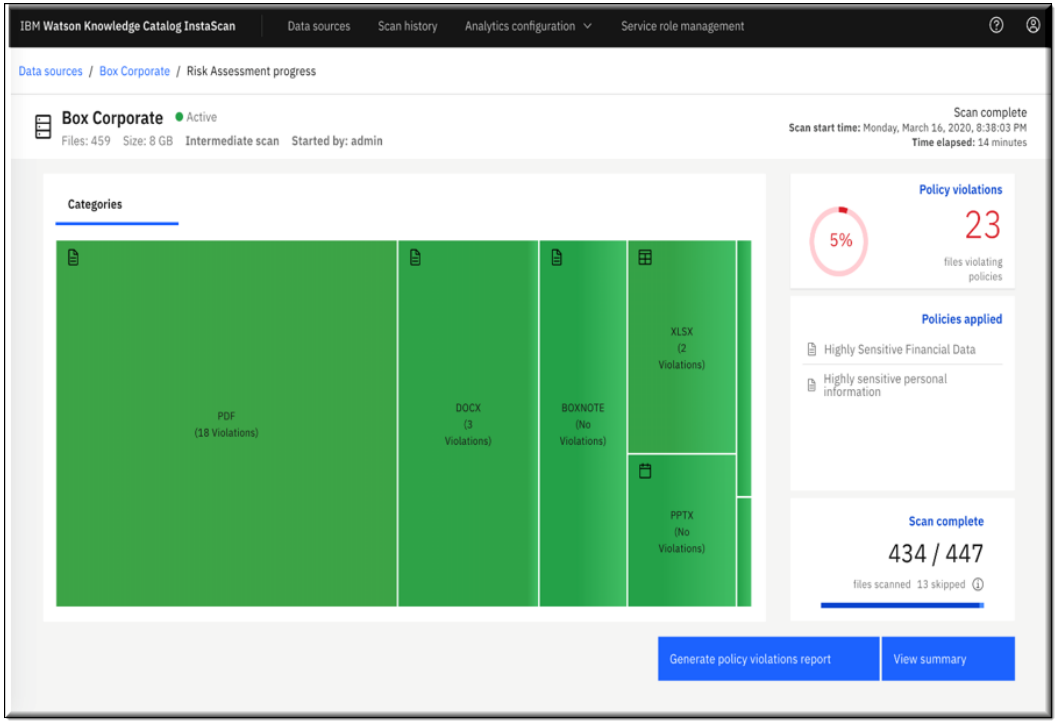
- In ungoverned catalogs for structured data assets
- In projects for both structured and unstructured data assets



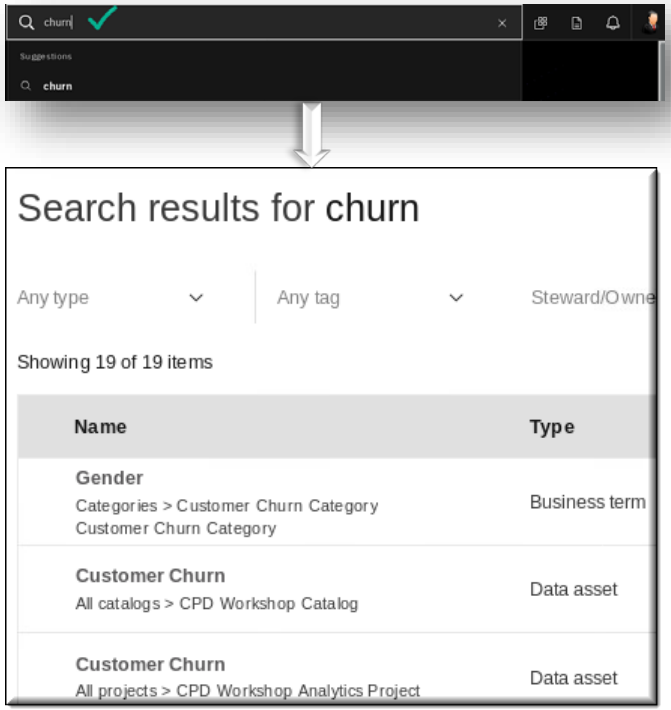
CPD Organize
InstaScan

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

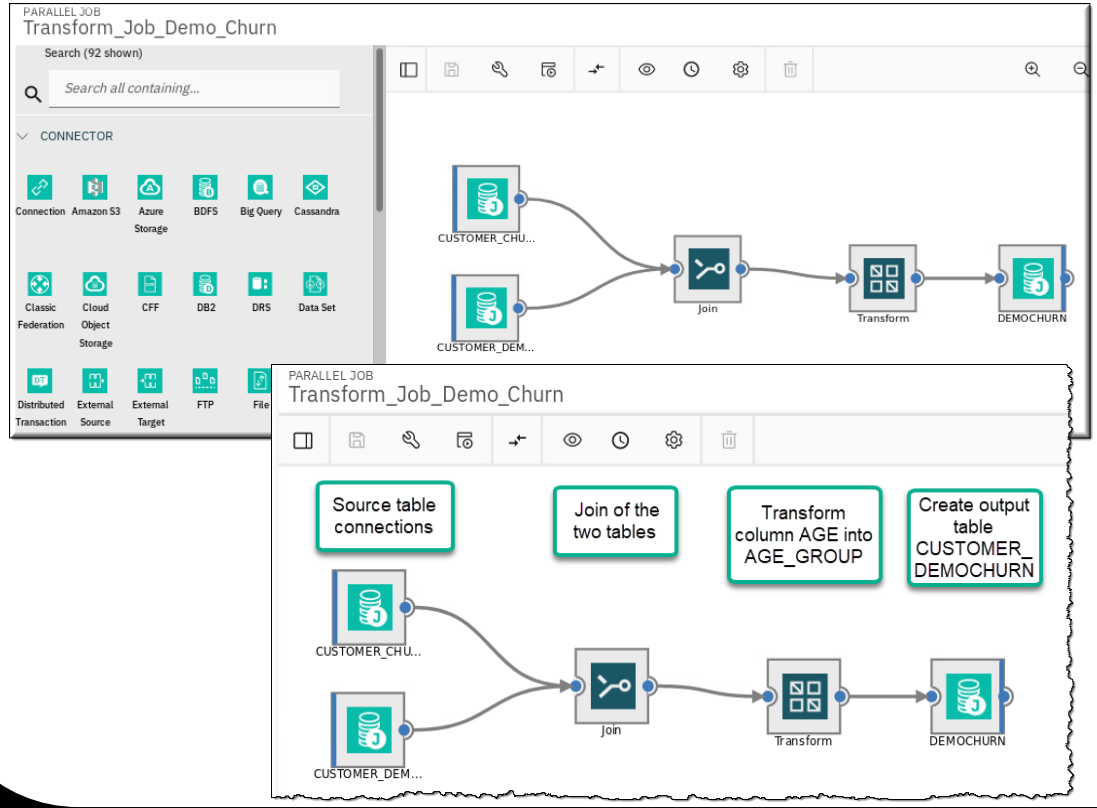


Relevancy of search results factors:

Text match	The provided text is searched in the asset name & description, where name contributes more to the higher place in the result list.
Asset rating	The higher the average rating the asset has, the higher it is on the results list.
Comments	The higher the number of comments, the higher the asset is on the results list.
Context match	The search results list might contain the closest neighbors of assets that are returned based on the text match.
Modification date	The assets that were modified recently are more likely to be returned in the search results.
Quality score	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.
Usage	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

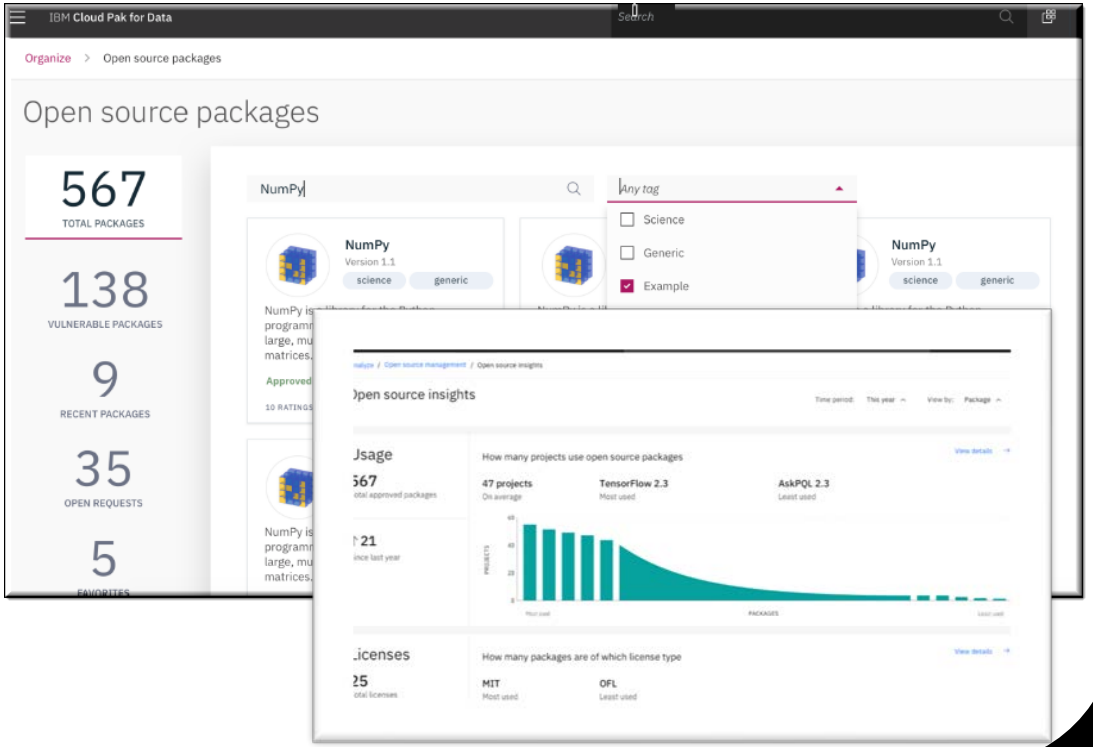


- Use powerful data transformation capabilities
- ML infused Smart Job Clustering, Smart Job Assist and automated job sequencing
- Design once, run on any-cloud using Kafka
- Remotely execute job for co-located access to data, satisfy geo-political requirements and save costs
- In-line data quality and 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



CPD Organize
Master Data Connect

Master Data Connect is an extension of:

- ✓ IBM Master Data Management Advanced Edition
- ✓ IBM Master Data Management Standard Edition

Services the majority of the MDM system workload - specifically reading MDM data.

Provides clients the opportunity to deploy a read-only instance of their master data entities on Cloud Pak for Data.

Majority (typically 80%-95%) of MDM workload is read/inquiry.

Mastered entities represent the highest quality data enterprises have about key domains like person, patient or organization.

Connect master data to your applications via RESTful APIs or IBM App Connect.

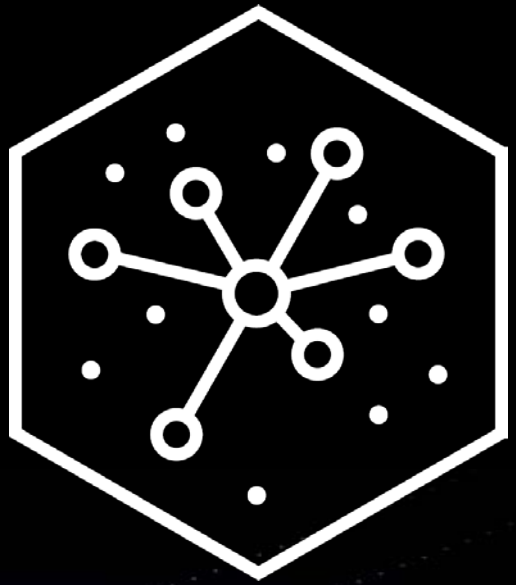
IBM Analytics Modernization Workshop

Part 2

• Introduction • Business Use Case	• Lab 01 • Lab 02
• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

IBM Analytics Modernization Workshop

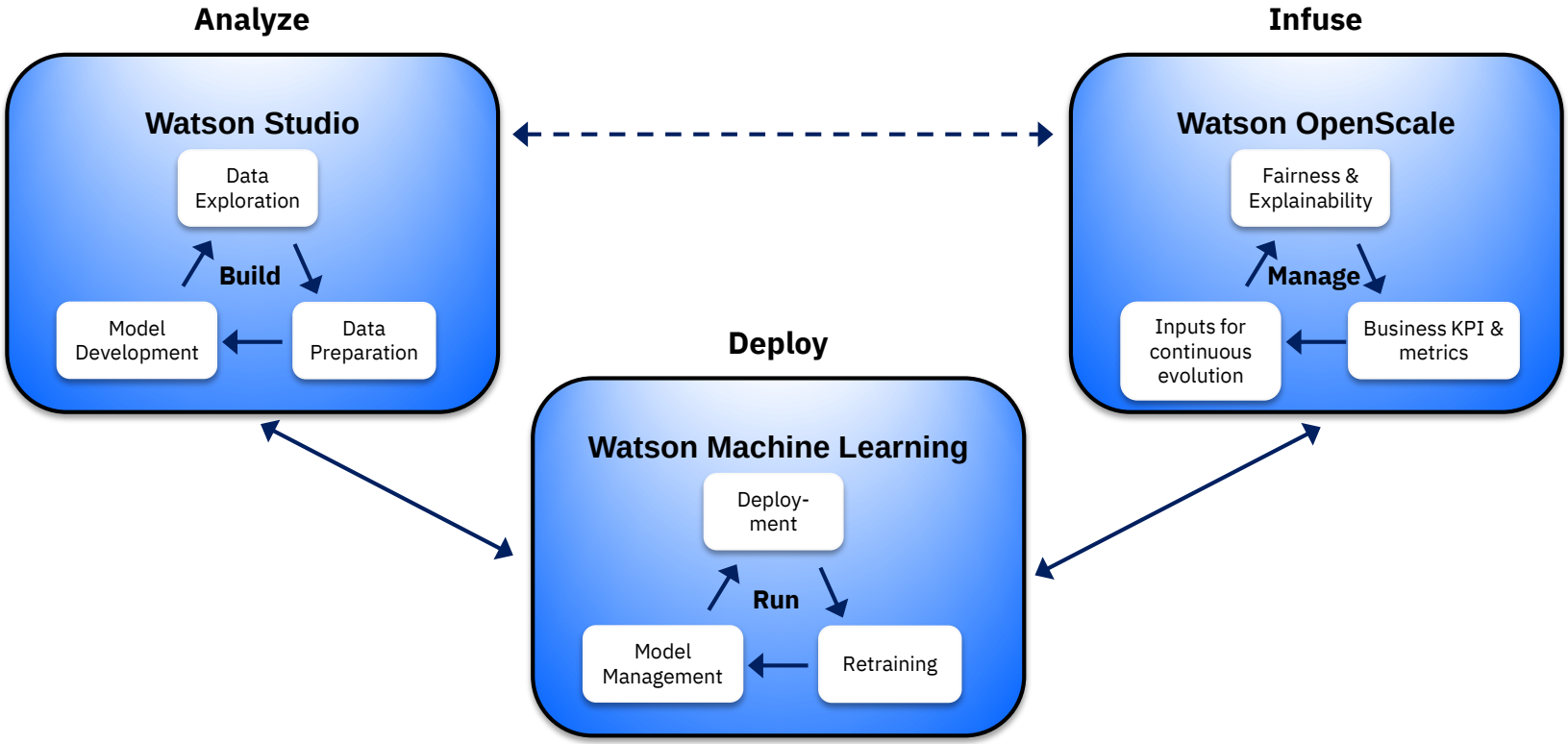
	• Introduction • Business Use Case	• Lab 01 • Lab 02
	• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
Part 3	• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10



Analyze

Lab 06 – Analyze: AutoAI & Notebooks

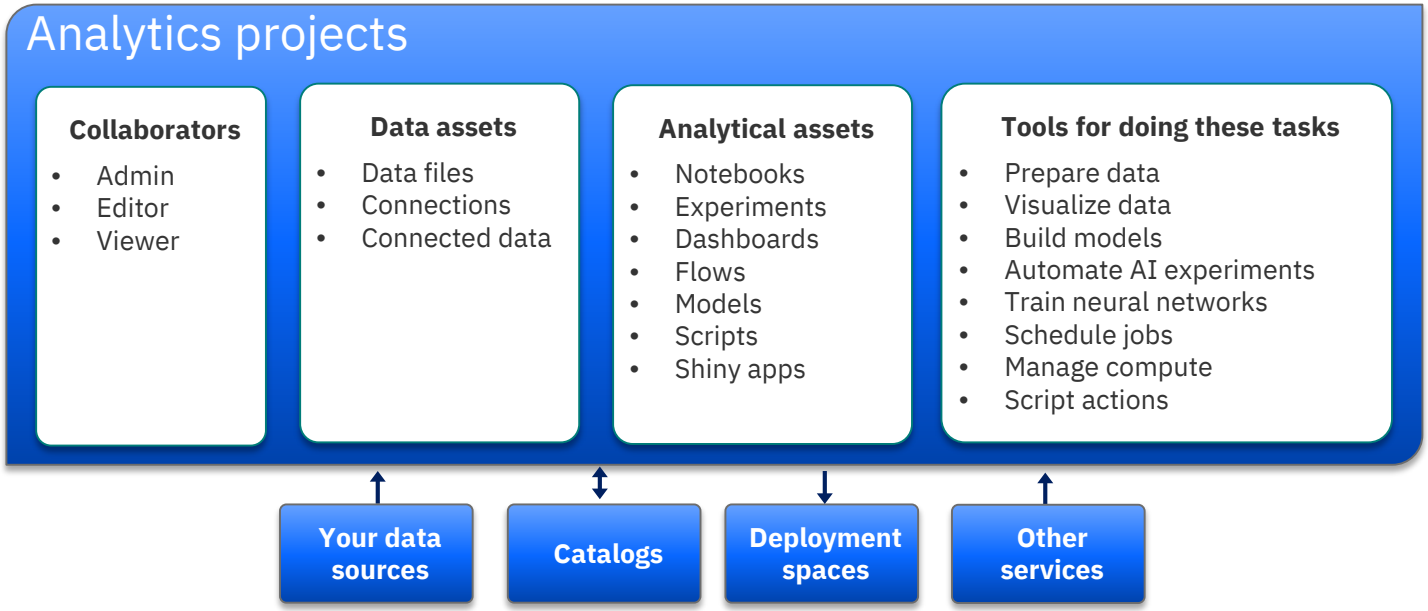
Analyze
The Data Science Lifecycle: Overview



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

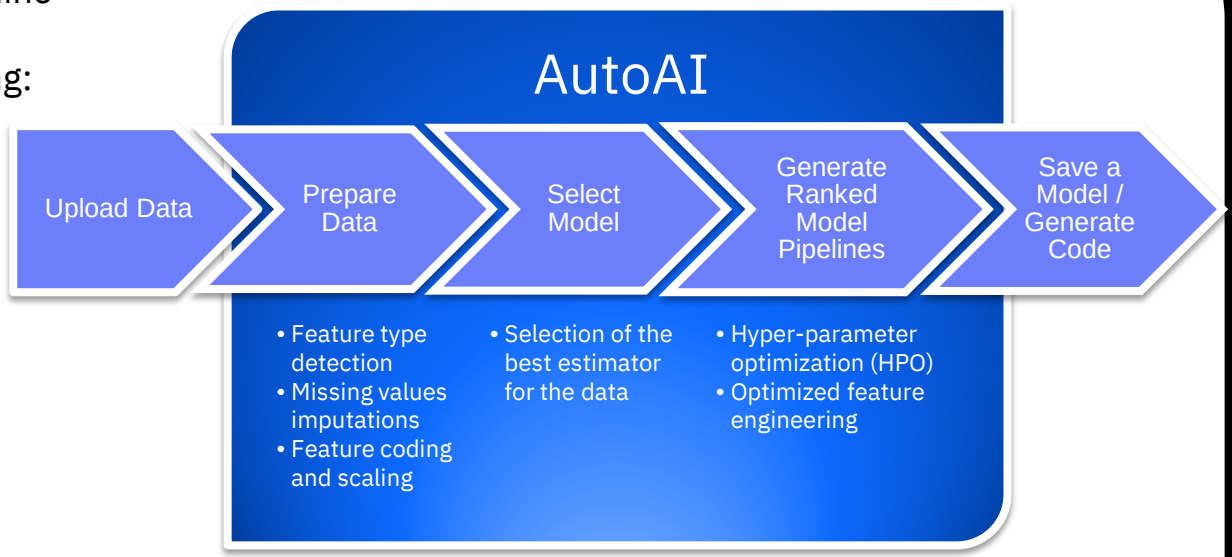
Analyze
AutoAI * – Overview

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**

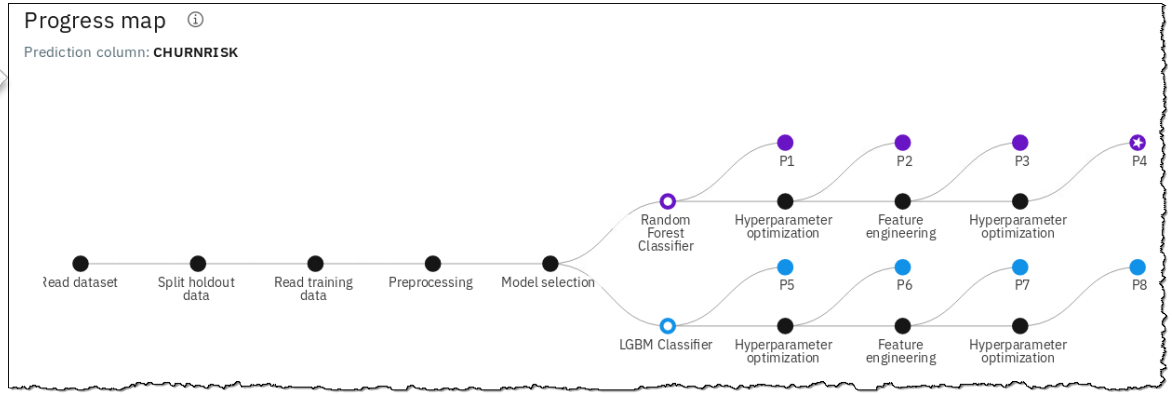


* AutoAI is enabled with the Watson Machine Learning service install, but it is driven through a Watson Studio Analytics Project

Analyze
AutoAI – Infographics

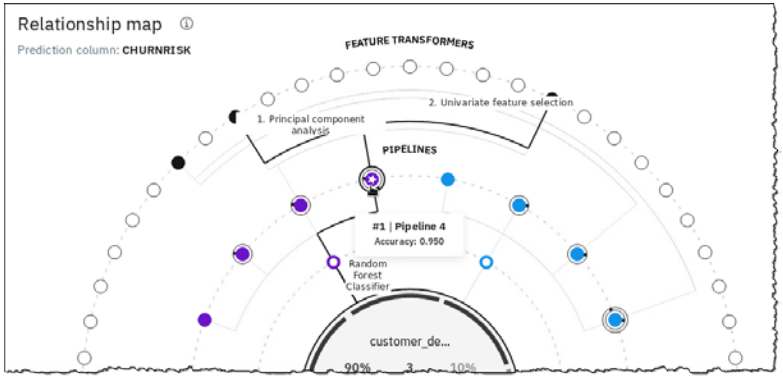
AutoAI Progress map

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

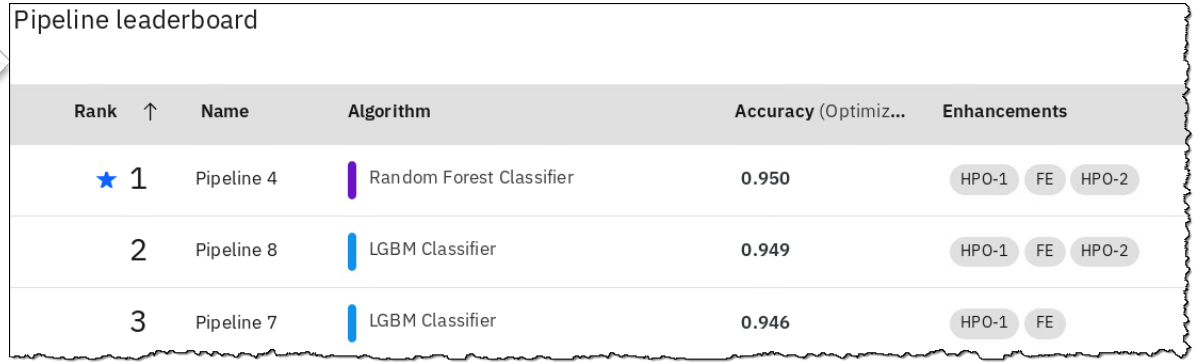
AutoAI pipeline leaderboard

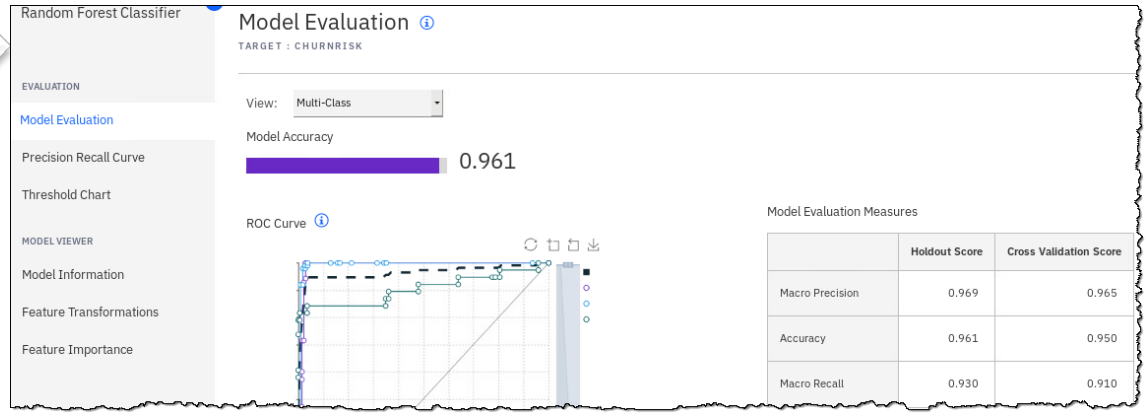
Shows the ranking of the pipelines for each potential model, the higher the better.

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**



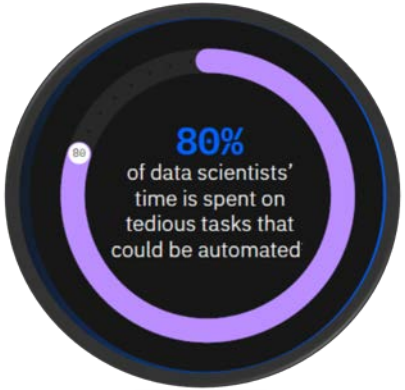


Analyze
AutoAI – Benefits

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

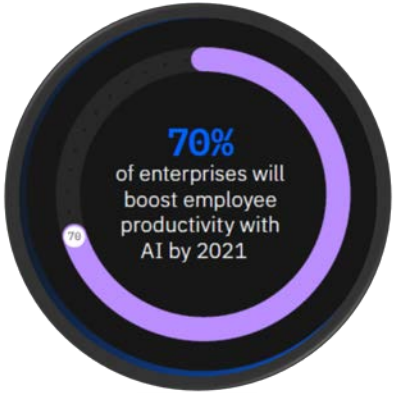
Go live with better models using the skill sets you have

Increase repeatability and minimize human intervention

3 Drive Productivity

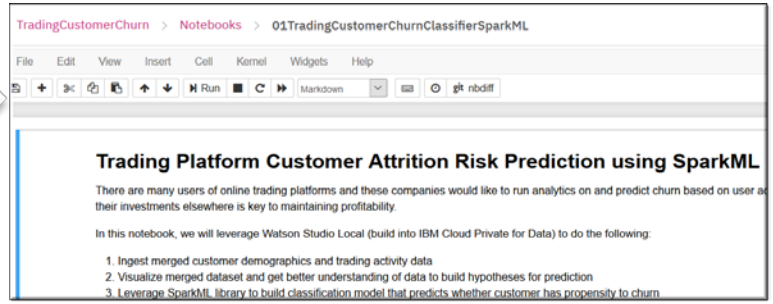
Get started with AI experiments without knowing how to code

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

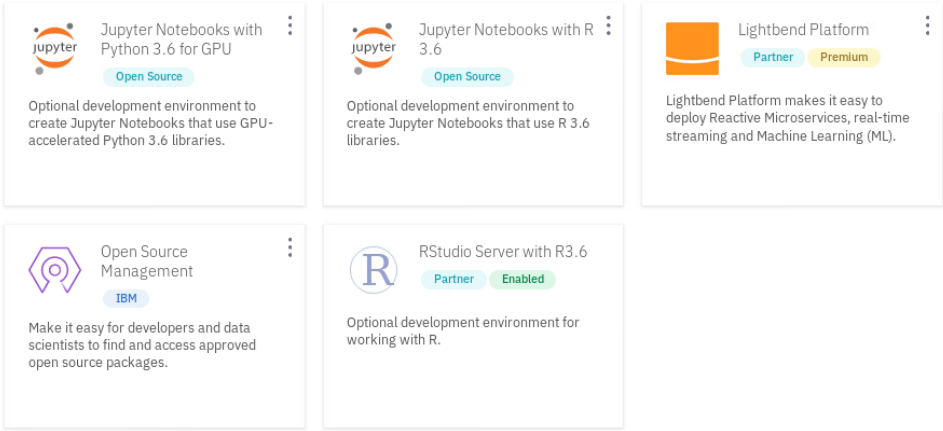


Analyze
Notebooks, RStudio and other tools

The default notebook environment:
Jupyter Notebook with Python 3.6

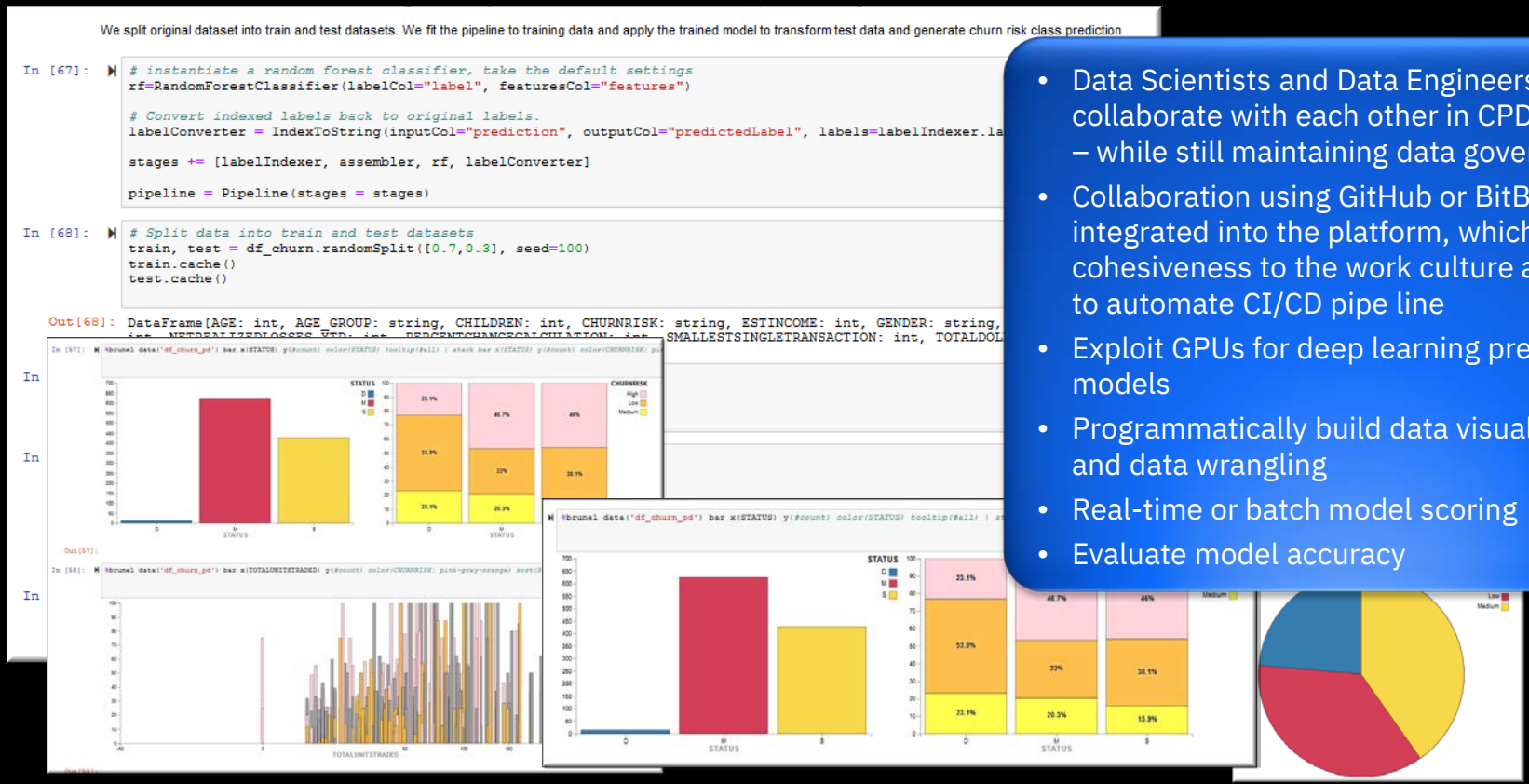


Developer tool services available:
- Jupyter Notebooks with Python 3.6 for GPU
- Jupyter Notebooks with R 3.6
- RStudio Server with R3.6
- Lightbend Platform
- OpenSource Management



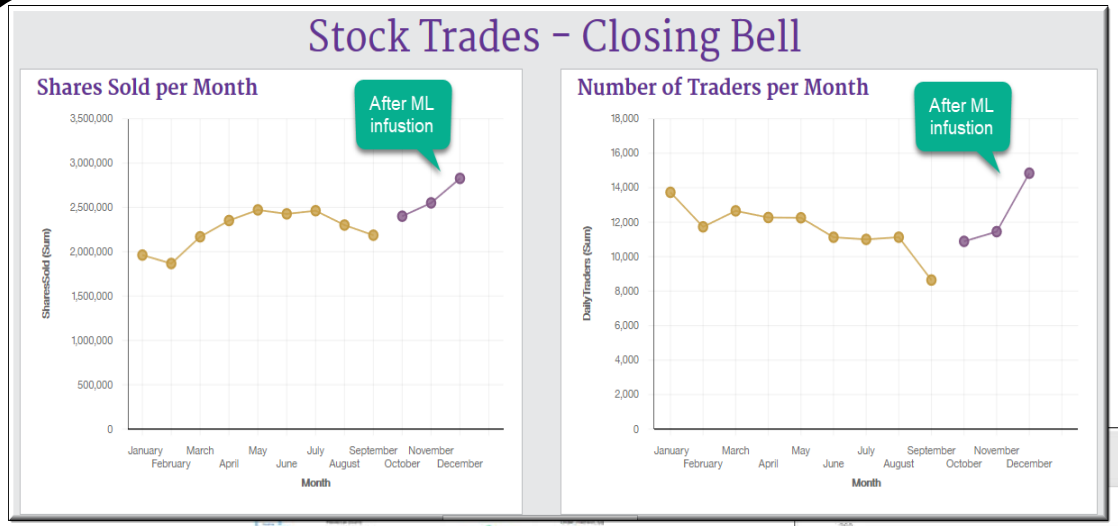
Analyze

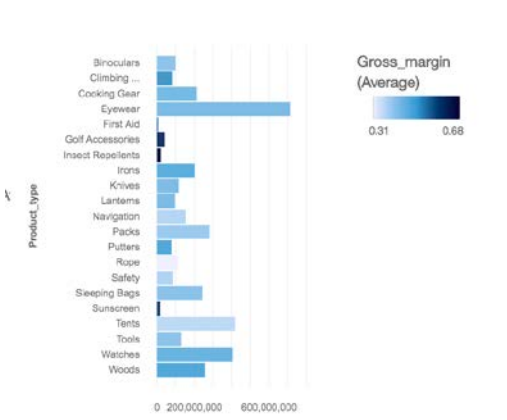
Watson Studio notebooks: Build Data Science & Machine Learning models

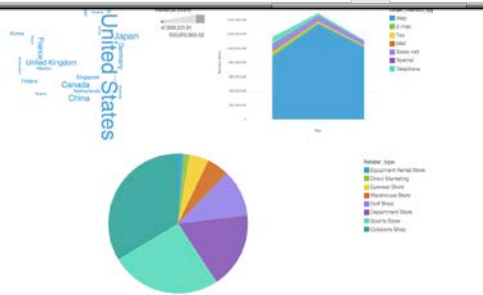


- Data Scientists and Data Engineers collaborate with each other in CPD platform – while still maintaining data governance
- Collaboration using GitHub or BitBucket is integrated into the platform, which brings a cohesiveness to the work culture and helps to automate CI/CD pipe line
- Exploit GPUs for 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

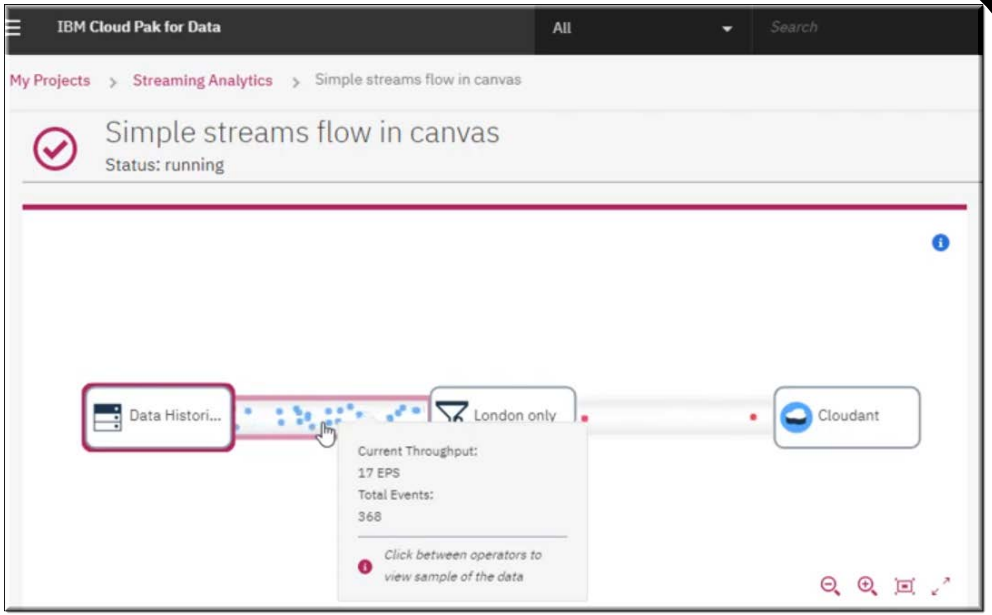
- No business disruption—run, score & update models continuously
- Machine learning, natural language, spatial-temporal, acoustic time series, etc.

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.

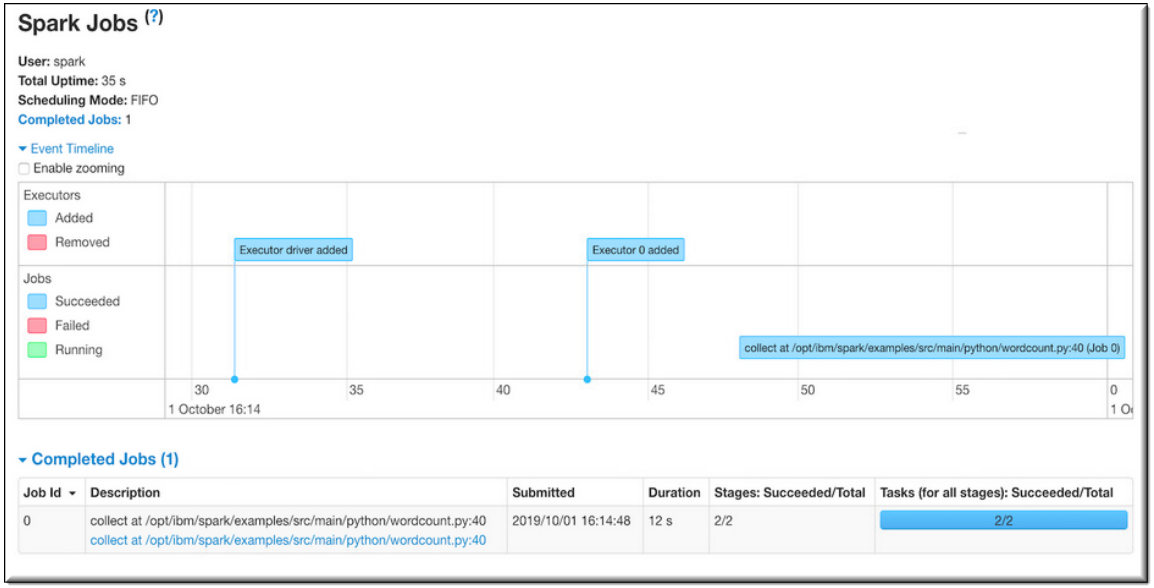
Operator types include: **sources, targets, data processing, alerts and real-time analytics**

Analyze
Analytics Engine - powered by Apache Spark

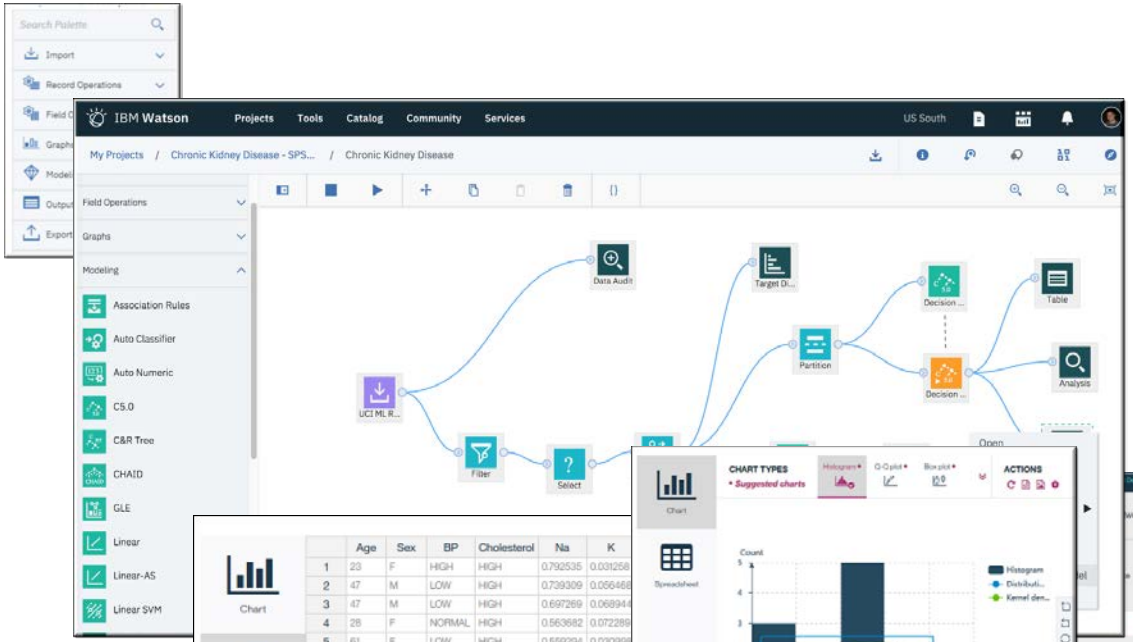
Analytics Engine is a serverless, performant, customizable and dedicated Spark engine that is available in seconds.

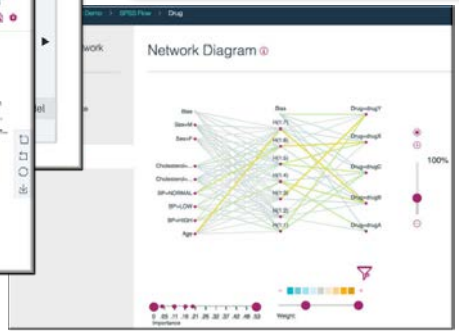
100% open source, Analytics Engine can run a variety of workloads on the CPD cluster:

- **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





SPSS Modeler

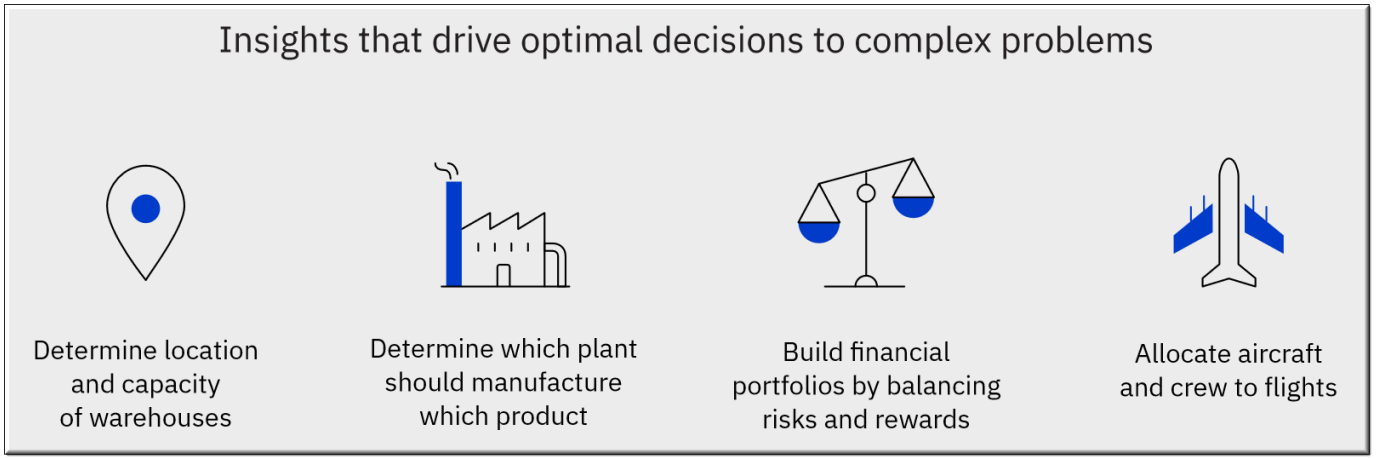
- A leading visual data science and machine-learning and predictive analytics solution
- Helps enterprises accelerate time to value and achieve desired outcomes by speeding up operational tasks for data scientists and business analysts
- Tap into data assets and modern applications, with complete algorithms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



Analyze
Watson Machine Learning : WML Accelerator for Deep Learning

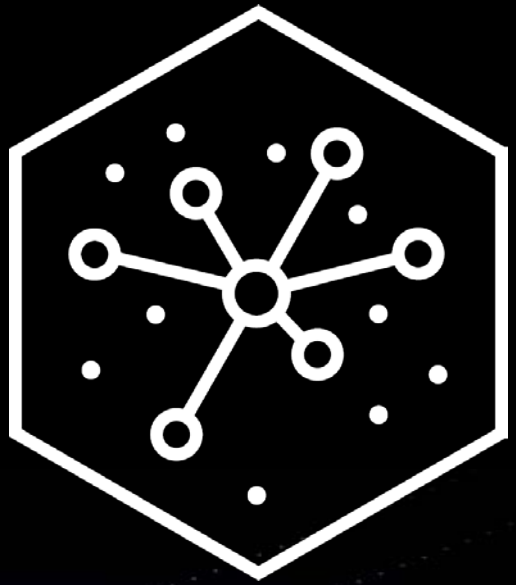
The **Experiment Builder** GUI interface is available from a WML project when you install the **WML Accelerator**. It is the simplest method to perform Deep Learning experiments.

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



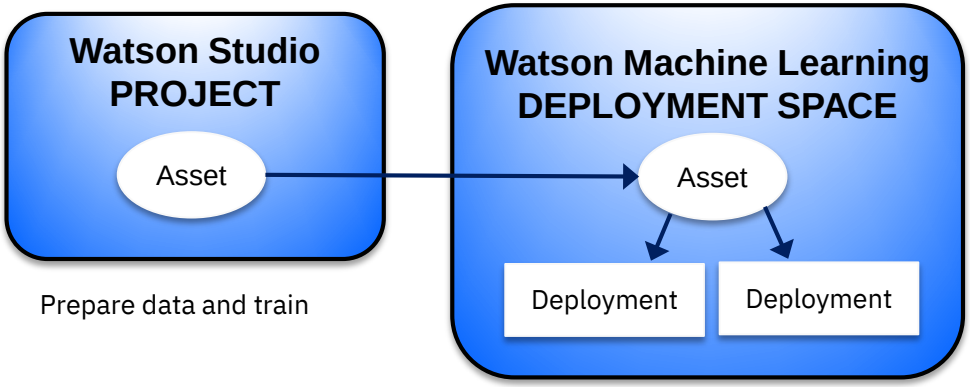
Deploy

Lab 07 – Deploy

Deploy
Watson Machine Learning: Deployment Spaces

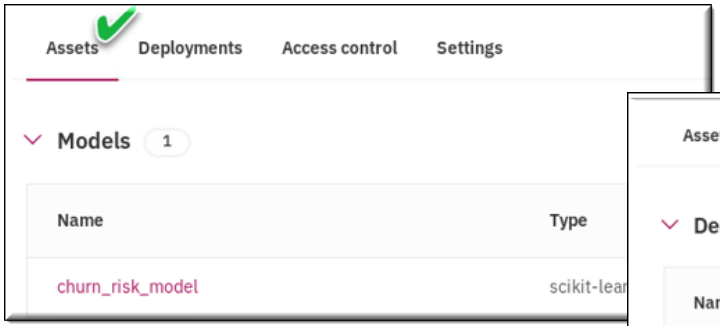
A **Deployment Space** is where you can:

- **Promote and save models**
- **Create the deployments from the models**
- **Find the information you need to score the model and get a prediction back**
- **Embed the deployment in an app so you can interact with it programmatically**



Prepare data and train

Configure, test, deploy and monitor



Deploy
Watson Machine Learning: Deployments

A **Deployment** is the last stage of the model development work. It means you put the model into production so that you can pass data to the model and return a score (or prediction).

After deploying a model, you can access the model *endpoint*, which you will need to make the model available for wider use in applications.

There are three type of WML deployments:

- **Online –** Provides an API endpoint needed to access the deployment programmatically to use in an application. Code snippets are provided in a variety of programming languages that illustrate how to access the deployment.
- **Batch –** Processes input data from a file and writes the output to a file.
- **Virtual** – For models downloaded into WML from different frameworks other than those built on the CPD platform.

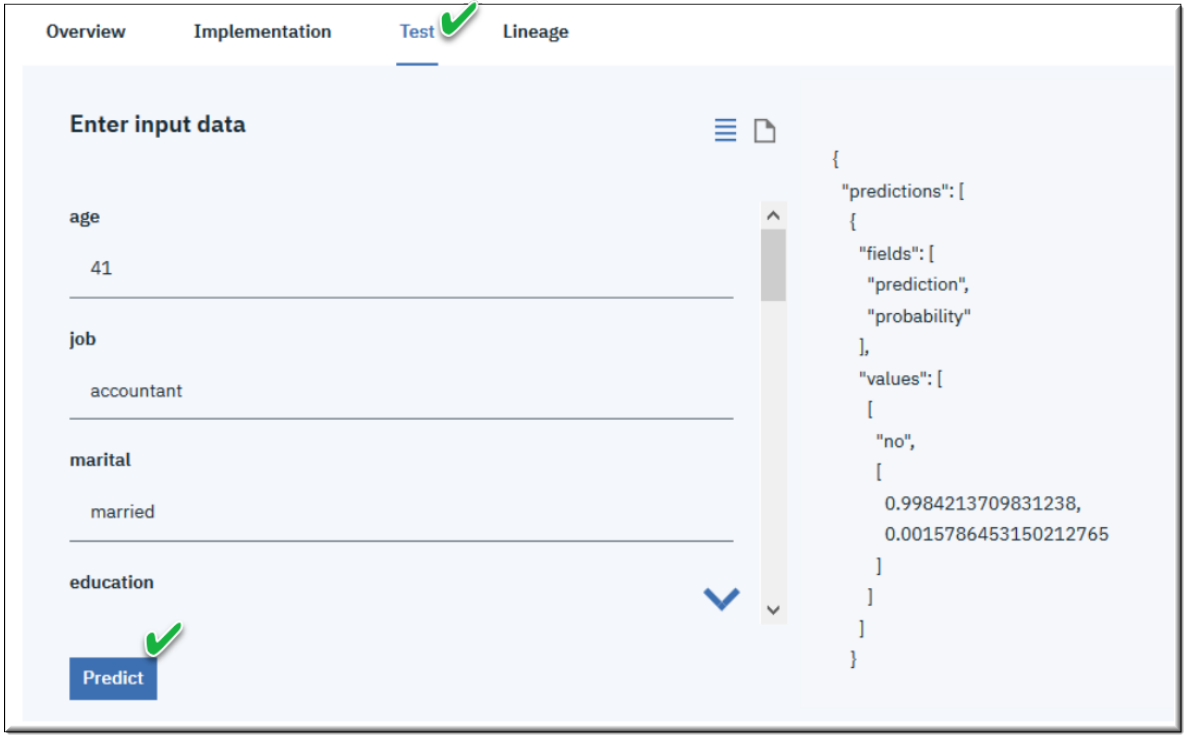
Deploy

Watson Machine Learning: Supported Frameworks for deployments

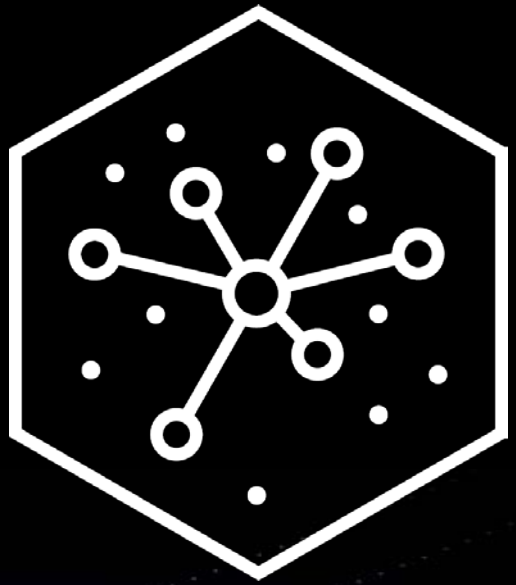
Framework	Version
Spark	2.3, 2.4
PMML	3.0 to 4.3
Hybrid/AutoML	0.1
SPSS	17.1, 18.1, 18.2
Scikit-learn	0.20, 0.22
XGBoost 0.82	xgboost_0.82
XGBoost 0.90	xgboost_0.90

Framework	Version
Tensorflow	1.15
Keras	2.2.5
Caffe	1.0
PyTorch	1.0, 1.1, 1.2
Decision Optimization	12.9, 12.10
Python Functions	0.1
Python Scripts	1.0

Deploy
Watson Machine Learning: Online deployment testing



- You can use the *form* in the *Deployment Space* to easily test an online deployment.
- On the *Test* tab of the Deployment details page, enter test data, and click *Predict* to see the result.



Infuse

Lab 08 – Infuse: Watson OpenScale
Lab 09 – Infuse: Cognos Analytics

Infuse
Watson OpenScale: Overview

Watson OpenScale:

- Automates and operates AI at scale across its entire lifecycle
- Delivers transparent, explainable outcomes freed from bias and drift
- Provides confidence in AI outcomes and spans the gap between the teams that operate AI and the business units that use these applications
- Monitors models developed in a 3rd party IDE, open source framework and hosted in a 3rd party or private model serve engine

Manage AI at Scale

Watson OpenScale

- Operations Dashboard
- Fairness & Bias Mitigation
- Drift Detection
- Explainability
- Business KPIs
- Payload Logging
- Data Mart

Model build / train frameworks

TensorFlow, scikit learn, Keras, PYTORCH, Caffe2, Spark MLlib

Model serving environments

kubernetes, Azure ML, aws

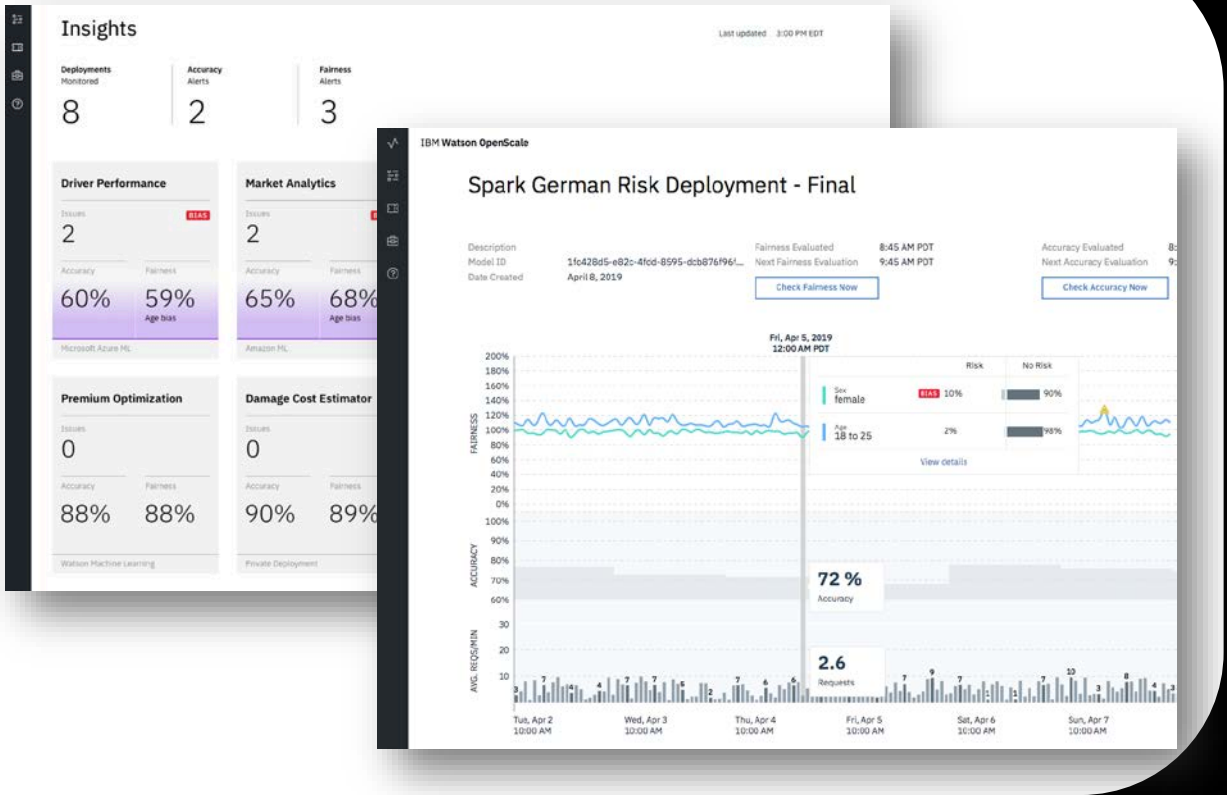
Infuse
Watson OpenScale: Operations dashboard

Description:

Monitor deployed models in a single dashboard that can be filtered by deployment making it easy to manage AI in apps

Value:

- Configure alerts or actions to be triggered when KPIs exceed threshold, ensuring model quality for improve business outcomes
- Measure model accuracy as it pertains to its ability to deliver outcomes more accurate than knowledge workers
- Provides “continuous evolution” for your models



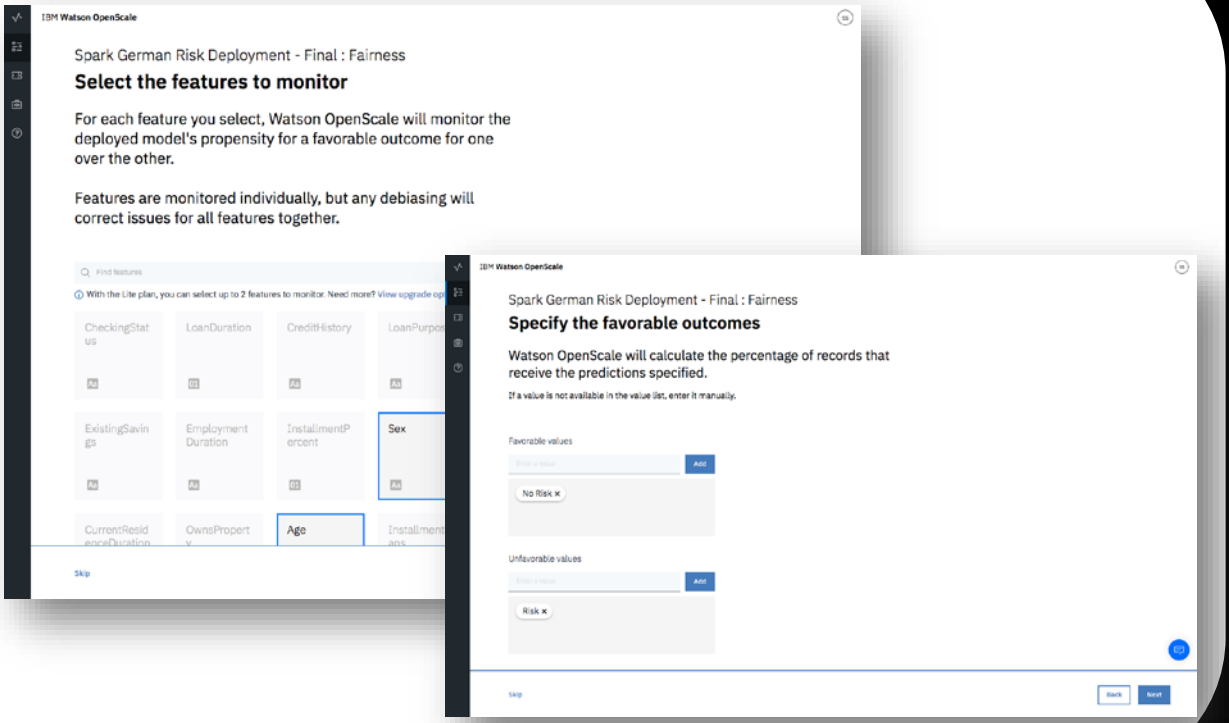
Infuse
Watson OpenScale: Model Fairness

Description:

Production Models need to make fair decisions and *can not be biased* in their recommendations

How it works:

- Outcomes are selected as “favorable or unfavorable”
- “Favored Populations” and “protected populations” are selected where majority and minority groups are found
- A score is calculated based on the probability of favorable outcome for minority vs. probability of favorable outcome for majority



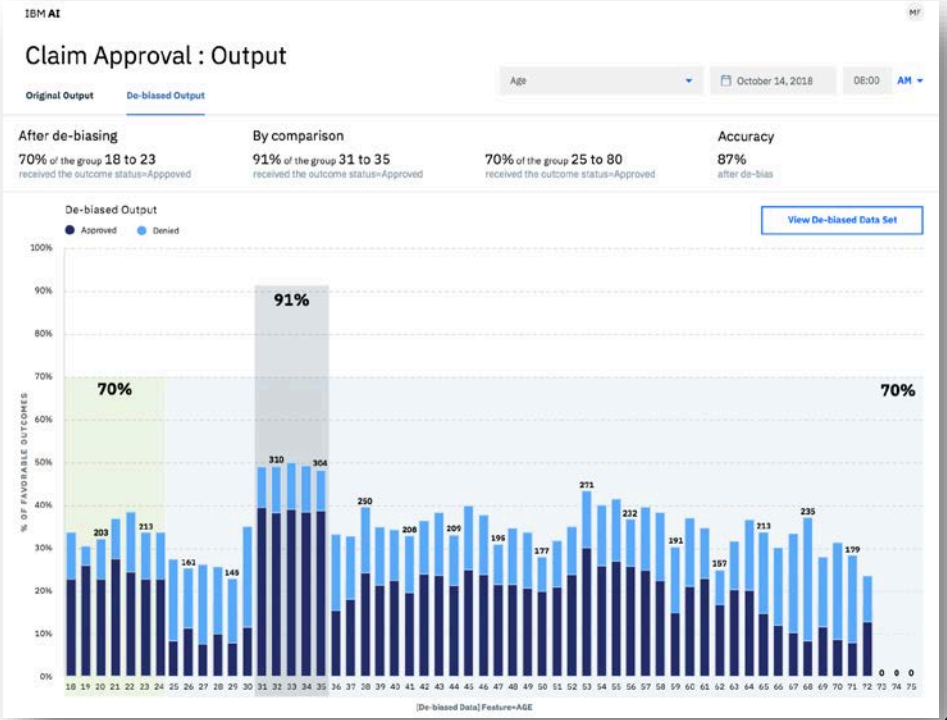
Infuse
Watson OpenScale: Bias Mitigation

Description:

Fairness is enforced with automatic bias mitigation.

How it works:

- Calculated on an *hourly basis* (over a sliding window defined by the user)
- Optimizations identify the *right subset of data to perturb* (rather than perturbing all the data)
- *Perturbed data is sent to the deployed* model to determine effect of perturbations
- An internal bias detection model (logistic regression) is built using perturbed data that *classifies whether new prediction will be biased or not*
- Users receive both the *original prediction* plus the *internal model's classification* of whether the monitored model's prediction is biased or not

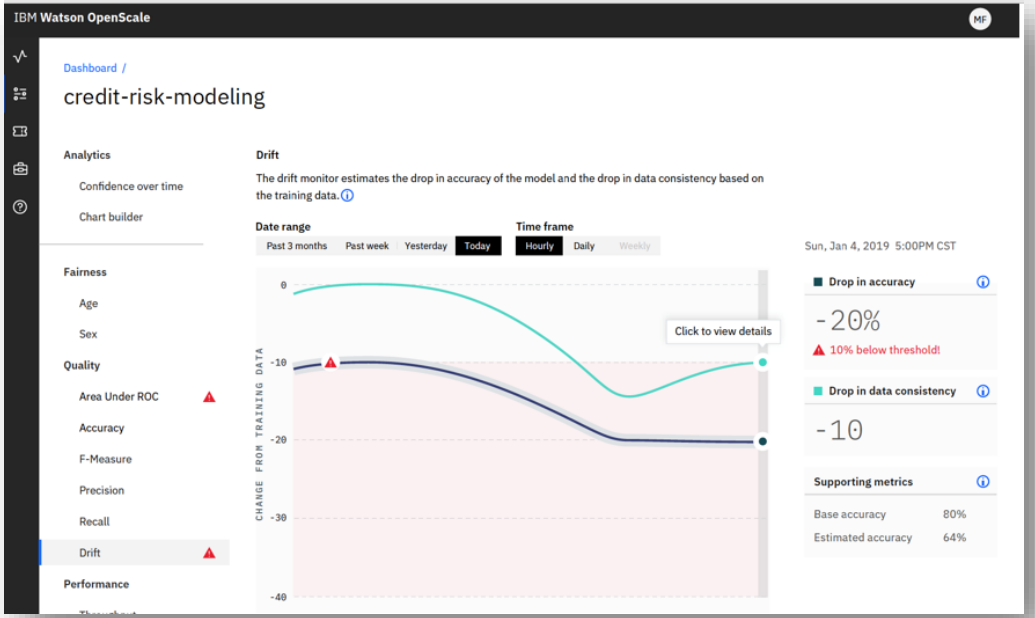


Infuse
Watson OpenScale: Drift detection

Description:

OpenScale monitors for two types of drift:

- **Drop in accuracy**: It estimates the drop in accuracy of the model at runtime. Accuracy could drop if there is an increase in transactions similar to those which the model was unable to evaluate correctly with the training data.
- **Drop in data consistency**: It estimates the drop in consistency of the data at runtime as compared to the characteristics of the data at training time.



OpenScale does drift detection on the entire payload data.

OpenScale measures the drift without requiring labeled data. Accuracy computation using labeled data can be expensive and might not be comprehensive

Infuse
Watson OpenScale: Explainability

Description:

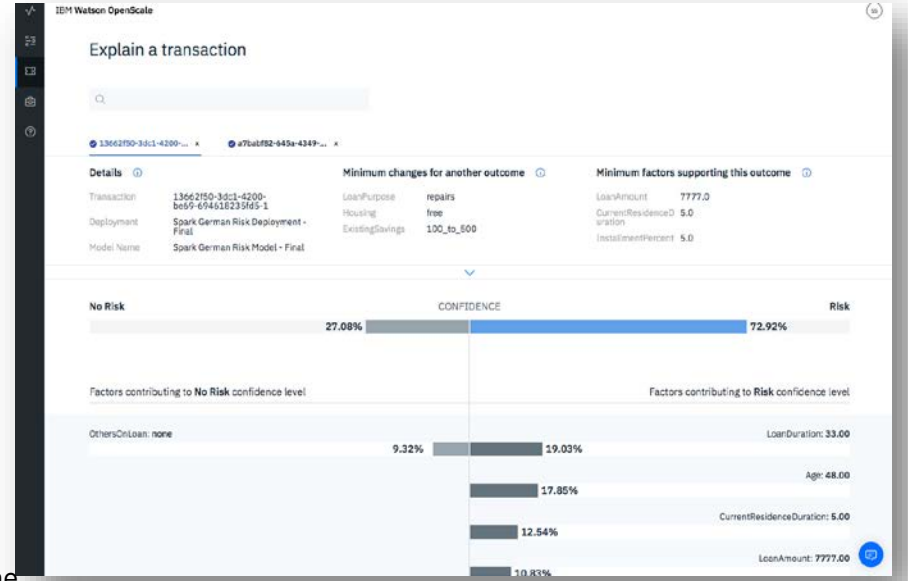
Allows you to understand which feature values of a model that are most influencing a prediction for a specific transaction

Example:

A loan is not approved by a model prediction - explainability will tell you why

How it works:

- Perturbation analysis on thousands of variations
- Risk model is created for two variations:
  - **LIME (local) Explanation:** set of features which played a positive or negative role in the prediction - also identifies the feature weights which helps to identify the most or least important features
  - **Contrastive Explanation:** Explains the behavior of the model in the vicinity of the data point whose explanation is being generated – assumption: the most common value is the least interesting from an explanation point of view

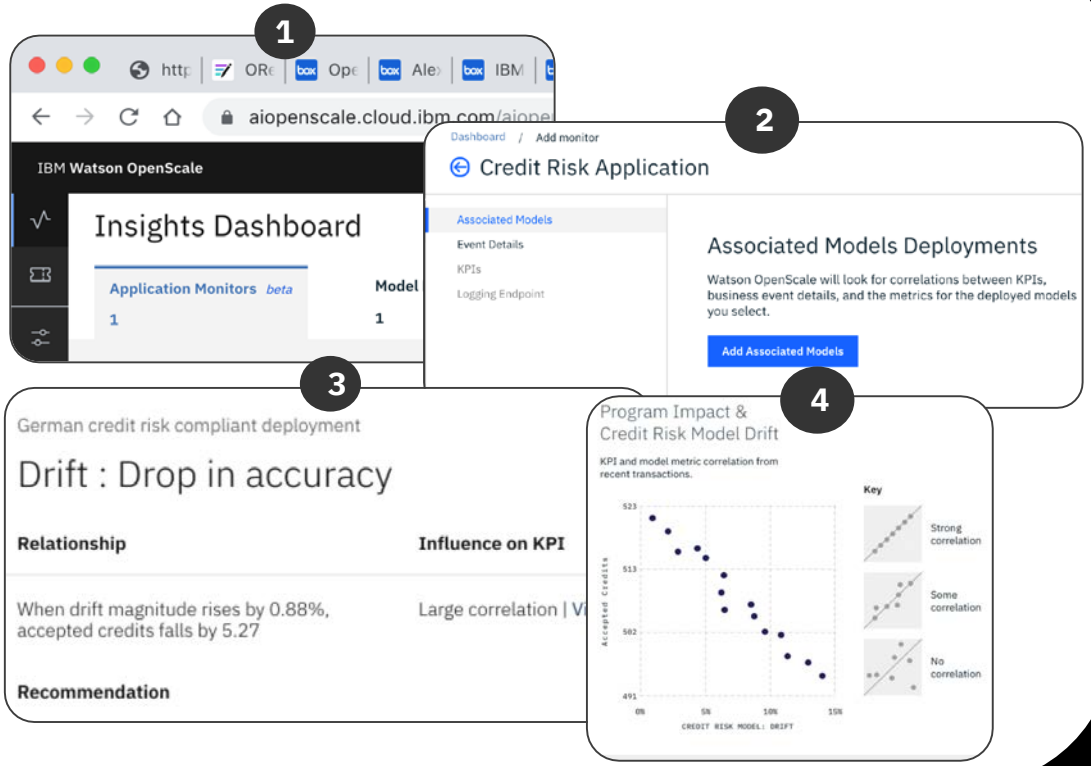


Infuse
Watson OpenScale: Business (Application) KPIs

Description:

The OpenScale UI dashboard contains an Application centric dashboard in addition to the Model centric dashboard:

- Bring in business event data into OpenScale and use it to compute Key Performance Indicators (KPIs)
- Monitor correlation between model monitors and KPIs - get alerts and recommendation from OpenScale
- Visualize the correlation through a plot



Infuse
Watson OpenScale: Payload logging

Description:

Payload logs capture (in Postgres) the request sent to the model or python function along with statistics about its health which when combined with feedback data provides insights into AI and application behavior

Value:

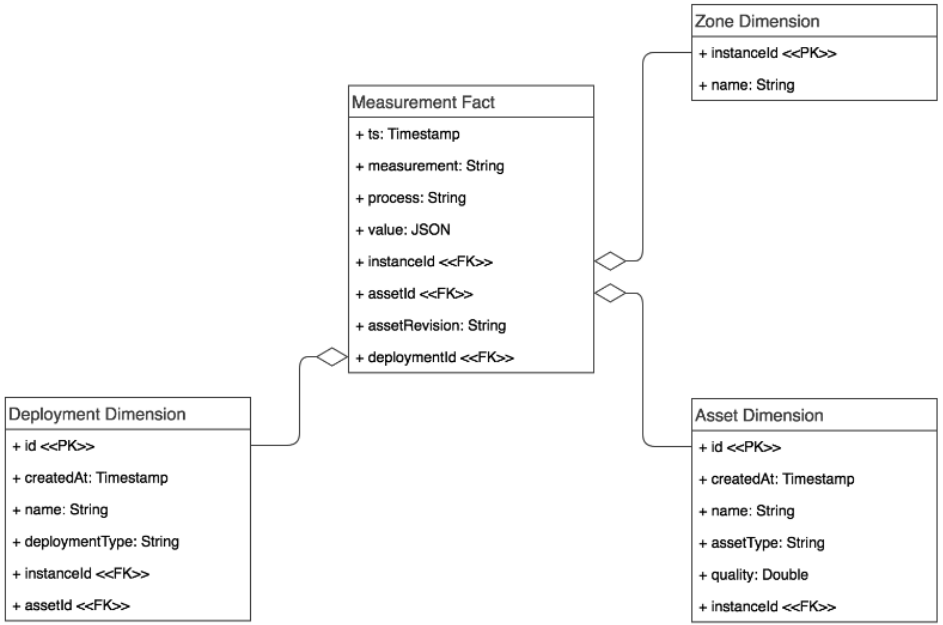
- Enable logging of payloads for *traceability* of business outcomes to AI recommendations
- Payload *data powers visualizations* for the OpenScale dashboard, making it easy to monitor health of deployed models
- Payload *fuels the Open Datamart* for custom reporting and business KPI integration

Infuse
Watson OpenScale: Data mart

Description:

The Data Mart is a data system to support the following use cases:

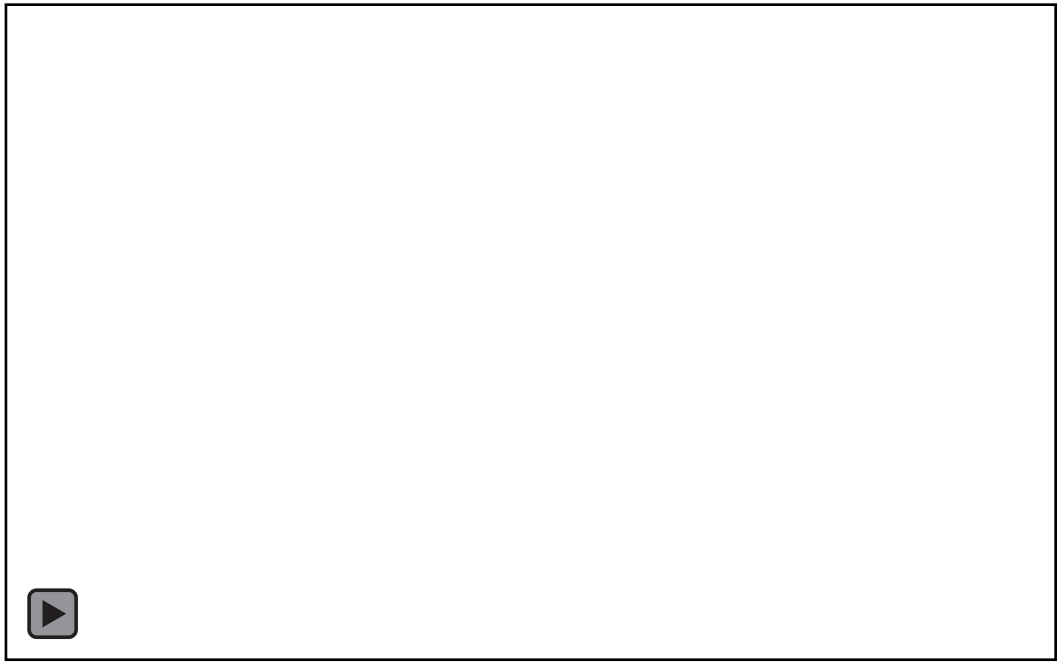
- Enable *OpenScale UI dashboards* for operations staff
- *Enable 3rd party reporting tools* so operations staff can customize their own dashboards using the underlying OpenScale data with their own tools
- *Enable data engineers to integrate OpenScale data* with existing enterprise data marts, warehouses, and data lakes by exporting data into those systems
- *Provide data scientists with the actual runtime payload, scoring and feedback data* that can be utilized in continuous learning



Infuse
Premium Service: Cognos Analytics

Cognos Analytics is self-service analytics, infused with AI and machine learning.

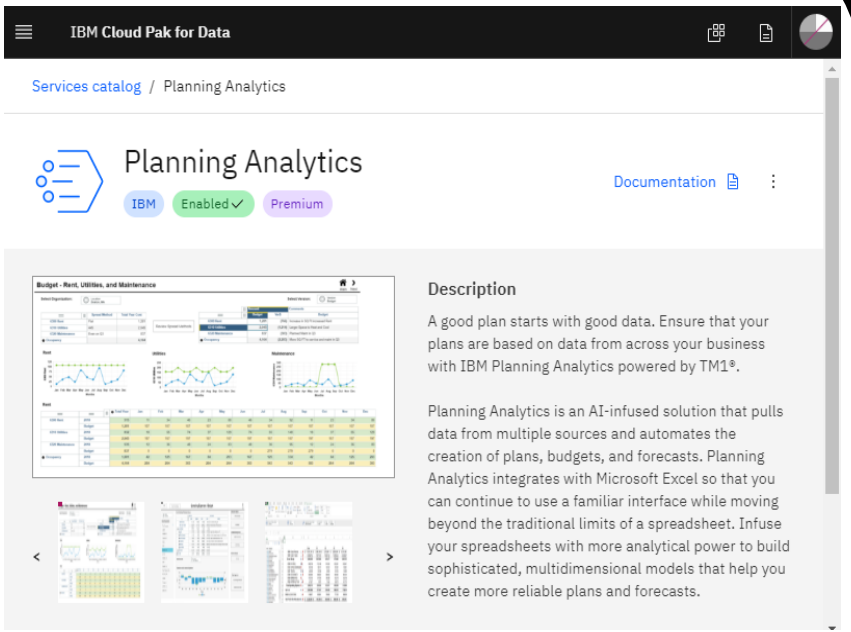
- Enables you to create stunning visualizations to share your findings through *dashboards* and *reports*
- These can be embedded (infused) into your applications
- The Cognos Analytics service makes it easier for you to extract meaning from your data with features such as:
  - **Automated data preparation**
  - **Automated modeling**
  - **Automated creation of visualizations and dashboards**
  - **Data exploration**



Infuse
Premium Service: Planning Analytics

Use Planning Analytics to:
- Automatically create plans, budgets and forecasts
- Steer business performance by bridging operations and finance for any department allowing you to adapt to changing business conditions
- See impact before executing – explore what-if scenarios and assess impact to determine the best course of action
- Make changes in real-time – pivot plans, budgets, and forecasts quickly to meet changing demands and priorities

Differentiators:
- Adjust financial plans in real time across departments
- Protect your investment in Microsoft Excel while transcending limitations of spreadsheets
- Uncover deep insights through AI-infused planning, without the need for help from a data scientist



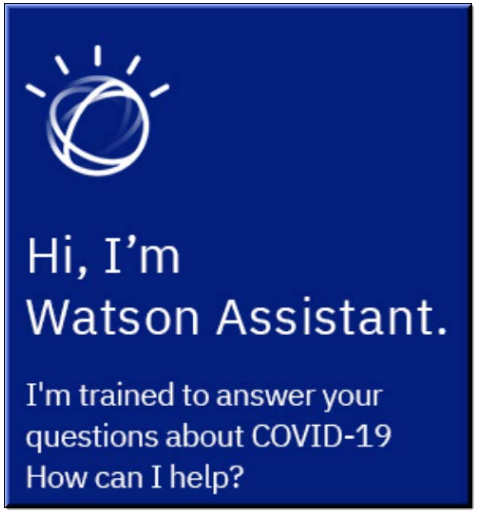
Infuse
Premium Service: Watson Assistant

Watson Assistant enables you to build conversational interfaces (chat bots) into any application, device, or channel

- Intuitive tooling for dialog building
- Intent recommendations from chat logs for continuous improvements
- Digression handling when user changes topics mod-conversation
- Out-of-the-box integration with search capabilities when coupled with Watson Discovery
- Common misspellings are automatically handled by Watson Assistant.
- Contextual entities: Support for additional languages not yet in Cloud Pak
- Fuzzy matching: Support for additional languages not yet in Cloud Pak

Benefits:

- Get up and running quickly
- Easily improve customer experience
- Reduce costs and speed of resolution
- Increase value by integrating apps and channels
- Ensure security, resiliency and data privacy - anywhere



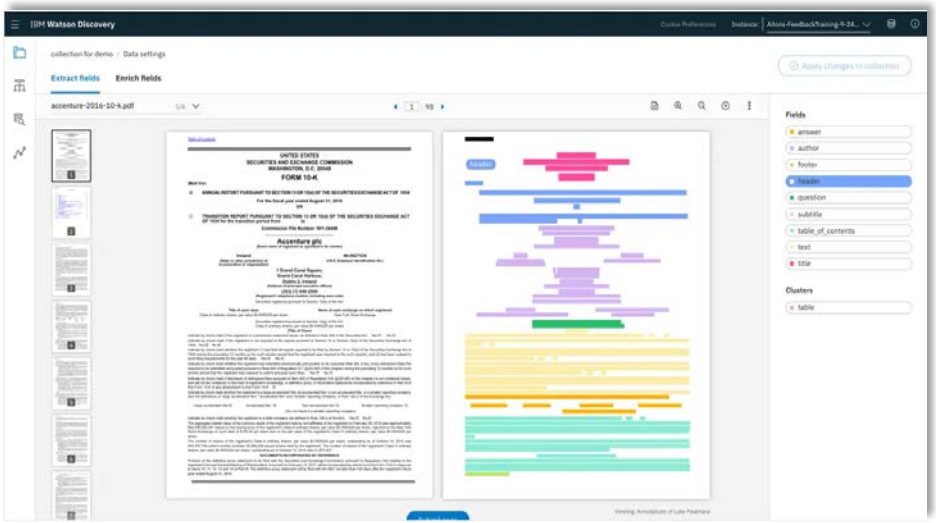
Infuse
Premium Service: Watson Discovery

Watson Discovery Surface answers and rich insights from your enterprise data.

- Award-winning enterprise search and AI search technology that breaks open data silos and retrieves specific answers to your questions while analyzing trends and relationships buried in enterprise data.
- Applies the latest breakthroughs in machine learning, including natural language processing capabilities, and is easily trained on the language of your domain.

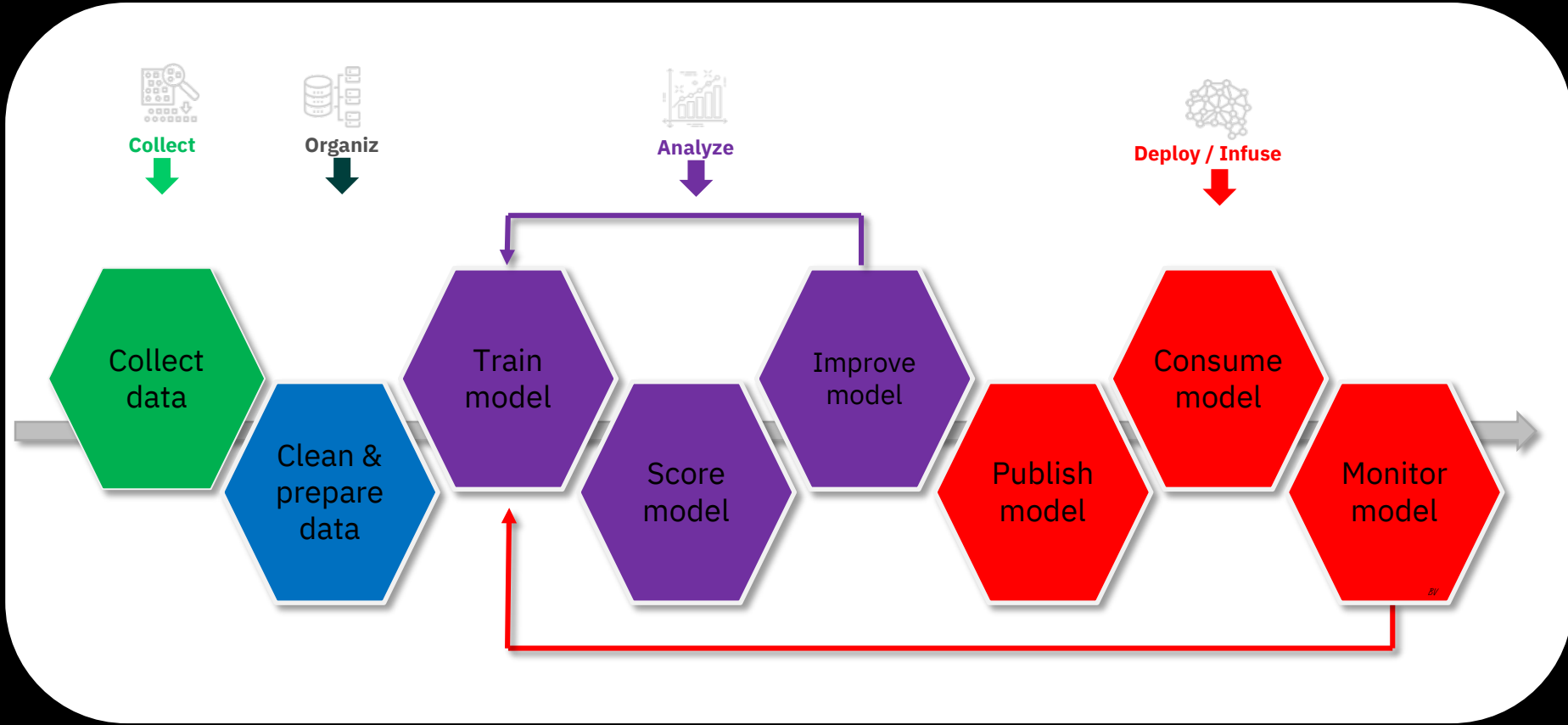
Benefits:

- Traditional enterprise search engines don’t provide exact answers because they can’t understand the nuances of phrases and acronyms in your industry and accurately search through your complex documents in a timely manner.
- Delivers specific answers to your queries while also serving up the entire document and supporting links, allowing your employees and customers to make informed decisions with confidence.



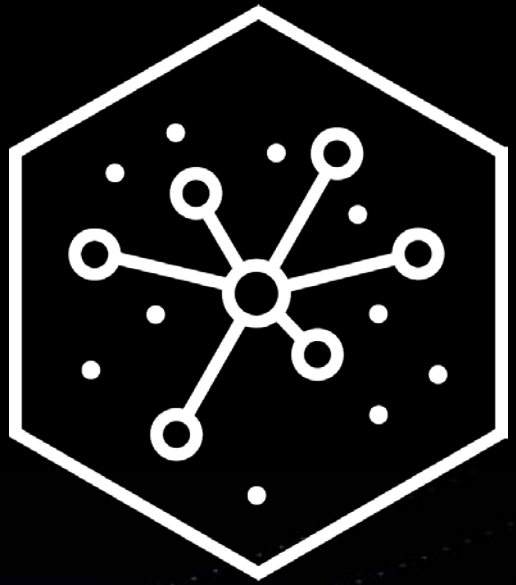
Machine Learning Model Lifecycle

CPD simplifies the entire process



Wrap up

Lab 10 – Wrap up



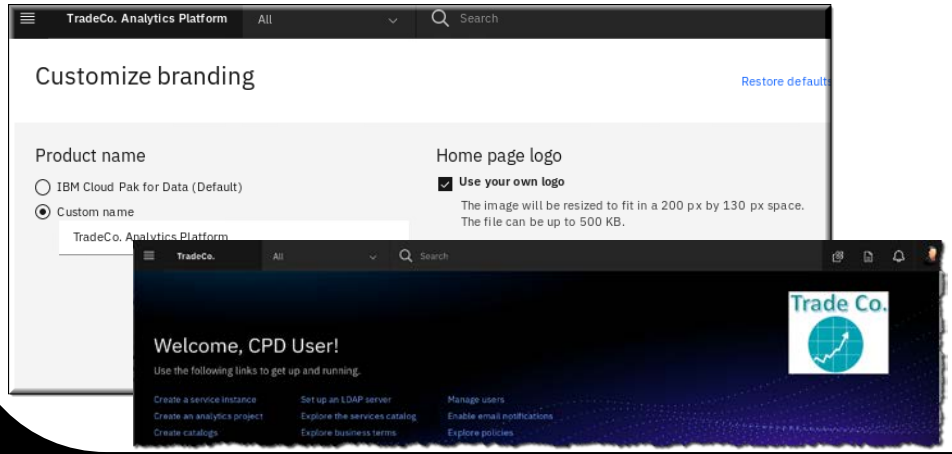
Cloud Pak for Data
Unified Experience

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 RHOCP 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: • Configure authentication • Configure platform • Manage users • Monitor platform
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

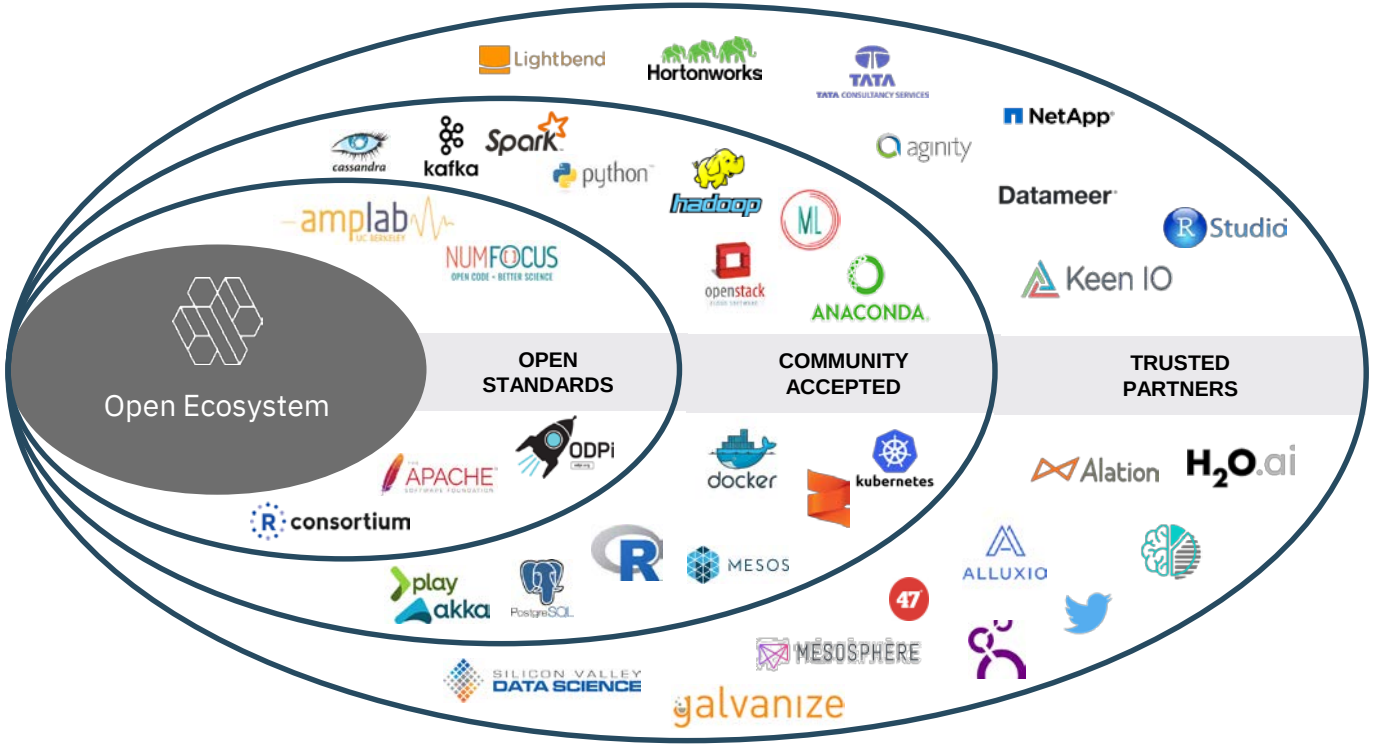
Compliance Best Practices:

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/

CPD Multi-cloud support

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

IBM Cloud

OPENSHIFT

openstack™

amazon web services™

Microsoft Azure

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

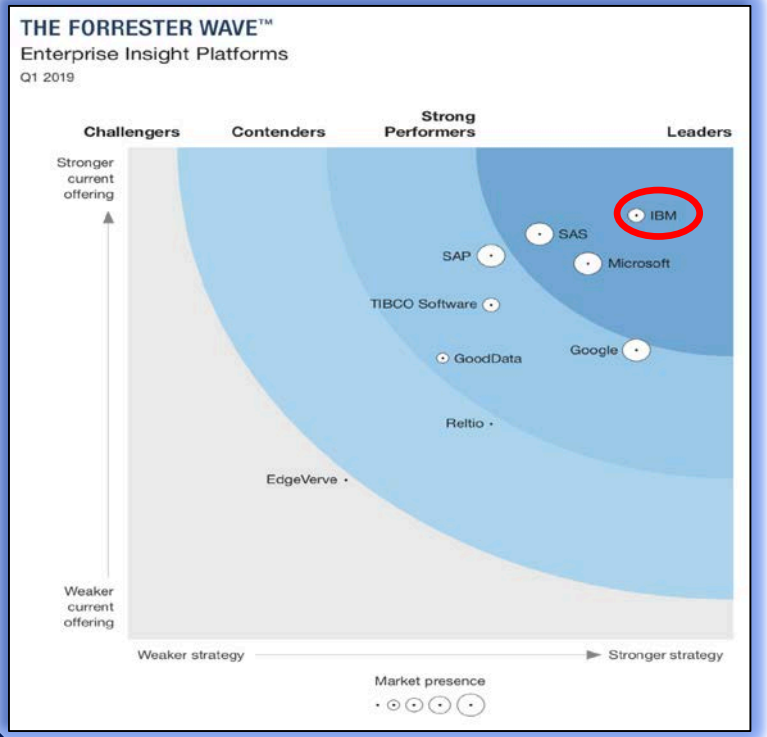
Over 24 Data Science & AI use-cases across 15 industry verticals (***applicability varies by customer***)

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 Services	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				
Clounter 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”



Report Preview : https://ibm.box.com/s/bry68nm9alduszvrffo105cvjhmde7pn

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”

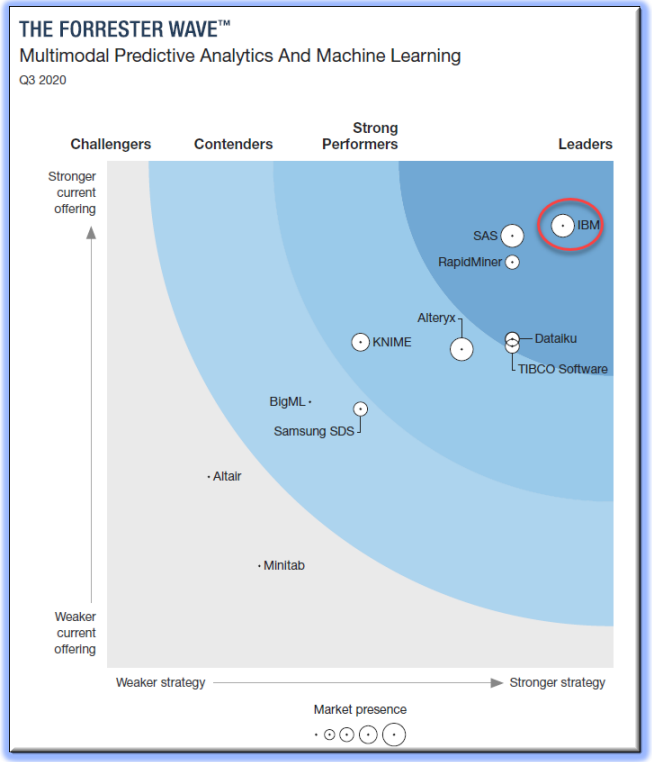
IBM has an impressive portfolio of individual data management and analytics capabilities that have consistently scored well on individual component Forrester Waves. With ICP for Data, IBM has pre-integrated capabilities that allow clients to be productive in a week or less. We were also impressed with 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.

Cloud Pak for Data

Ranked #1 by Forrester for “Multimodal Predictive Analytics And Machine Learning”



Report Preview : https://ibm.box.com/v/Forrester2020

Why IBM

- IBM is packed with AI lifecycle services everywhere — public, private, and on-prem. What do you get when you combine a full stack of data analytics capabilities — from data management to PAML to business intelligence — in a microservices framework that can run seamlessly on premises, in the private cloud, and in multiple public clouds?
- IBM Cloud Pak for Data, an offering that makes capabilities across the PAML lifecycle available when and where your users need them.
- The crown jewel of Cloud Pak for Data is Watson Studio, a PAML offering that combines easy-to-use, SPSS-inspired workflow capabilities with open source ML libraries and notebook-based interfaces.
- IBM continues to add innovations from IBM Research like fairness monitoring, bias mitigation, AutoML, and federated learning.
- IBM offers a compelling, scalable, increasingly integrated, and harmonized platform with differentiated capabilities that spans the entire PAML lifecycle and can be deployed anywhere across any cloud. Users will have to navigate some technical previews and quickly evolving features, but that will save them from otherwise having to stitch together a mismatch of proprietary solutions and open source software.

CPD Packaging
Edition Overview

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

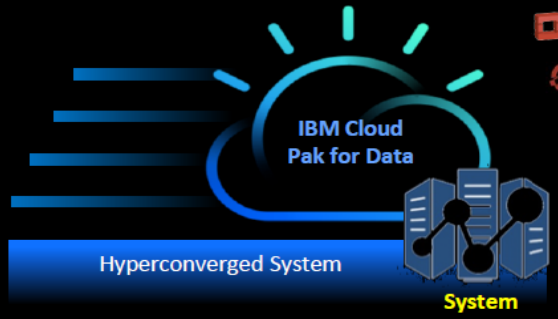
- Deploy in any form: on premises, on public cloud or CPD System
- Beginning w/ recommended minimum of 48 VPCs
- Select 24 VPC configurations supported
- Expands in increments of 1 VPC
- No maximum

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

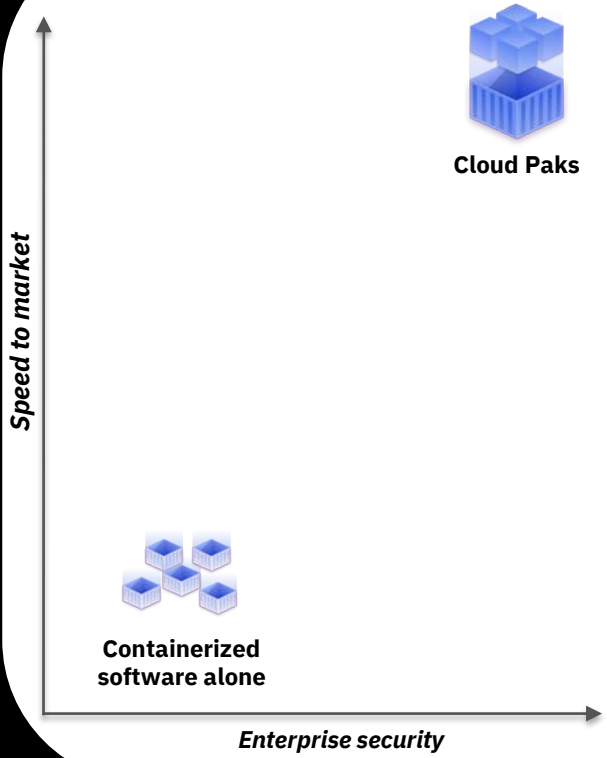
Deploy a complete private cloud in under 4 hours, with no assembly required

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



	Containers Alone Client creates containers or receives software as standalone container(s)	IBM Cloud Paks Complete solutions certified for enterprise use cases
Runs anywhere	Yes	Yes
Vulnerability scanned	Yes	Yes
Red Hat container certification	Depends on product	Yes
Complete solution w/ container platform	No	Yes
Flexible & modular	No	Yes
IBM certified/orchestrated for production (Built for Kubernetes by experts; certified against 200+ criteria)	No	Yes
Multicloud validation	No	Yes
Integrated deployment experience	No	Yes
Full stack support by IBM (Base OS, software, and container platform)	No	Yes
License metering integration	No	Yes
Scalable and resilient	No	Yes
Encrypted secrets / limited privileges	Do it yourself	Yes
Management and operations	Build your own	Yes
Lifecycle Management	Manage it yourself	Yes

Cloud Pak for Data
IBM Kubernetes Certified

http://ibm.biz/cp-certify

Production Grade

- Multi-cloud
- Storage
- Networking
- Resiliency
- Scalability
- Self Healing
- Recoverable

Security

- Vulnerability Management
- Limited Security Privilege
- Secure Access, Keys, Certs
- Network & Data Protection
- Secrets and SPI
- Security and Privacy by Design

Quality Assurance

- Comprehensive Testing
- Architectures
- OCP Versions
- Currency
- Airgap
- Upgrades

Lifecycle Management

- Patch
- Upgradeability
- Rollback/Restore
- Backup/Recovery
- OCP Version Alignment
- Kubernetes Currency

Consistency and Standards

- Consistent Packaging / Publishing
- Supporting Operators and Helm
- Consistent Entitlement management
- Common management of OSS elements
- UBI and Red Hat Certified
- Consistent use of OCP and IBM Services
- ~200 Code Standards enforced
- Governed Best Practice / Anti Practices

Red Hat OpenShift

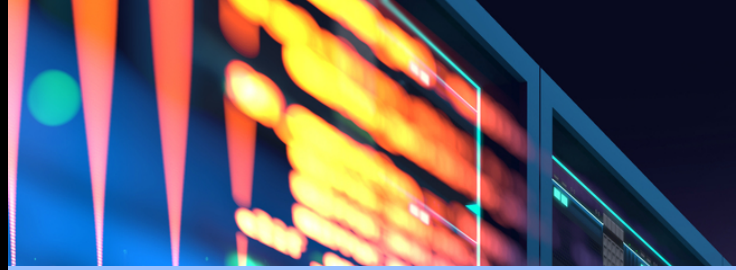
OPERATOR FRAMEWORK

redhat CERTIFIED

- Managed Image CVEs
- Packaging
- Publishing
- Trusted Source
- E2E Support

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*



Cloud Pak for Data

Industry: Telecommunications
Geography: North America

Turbocharged digital transformation

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.

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."

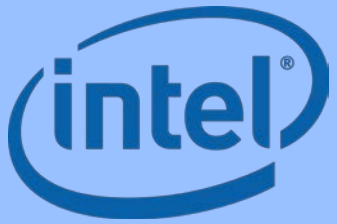
Michelle Gehl
VP Networks OSS Applications and Operations, Sprint

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)



Is AI your priority? Start with a data strategy

Read the blog and watch the video

Cloud Pak for Data

Industry: Technology
Geography: North America

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.

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

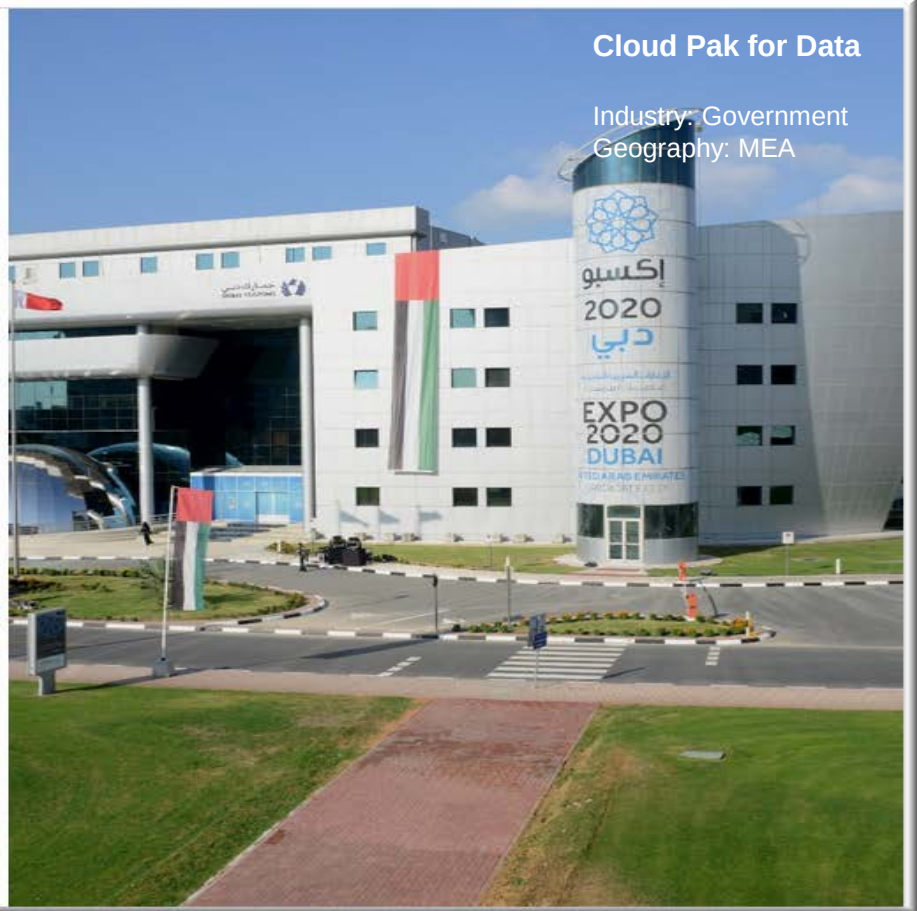
Senior Principal Engineer and Chief Data Scientist - Americas, Intel Corporation

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



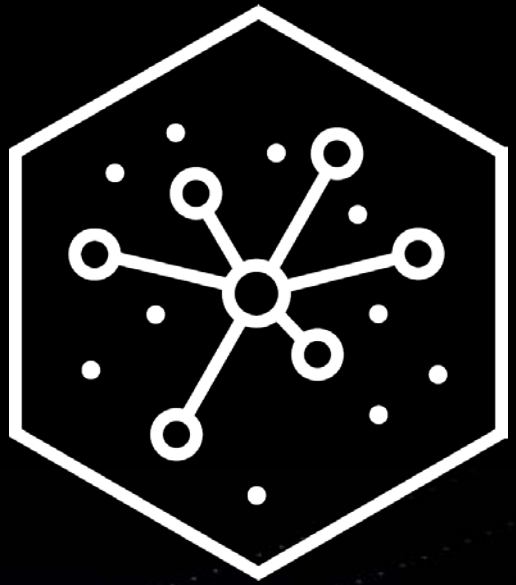
IBM Analytics Modernization Workshop

Part 3

• Introduction • Business Use Case	• Lab 01 • Lab 02
• Collect: Connect • Organize • Collect: Virtualize	• Lab 03 • Lab 04 • Lab 05
• Analyze • Deploy • Infuse – OpenScale • Infuse – Cognos Analytics • Wrap up	• Lab 06 • Lab 07 • Lab 08 • Lab 09 • Lab 10

Thank you for your time!

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



We appreciate your feedback.

Presentation by Burt Vialpando and Kent Rubin - IBM

Copyright and trademarks

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
July 2020