



IBM Journey to Cloud and AI Analytics Modernization Workshop

Featuring: Cloud Pak for Data

Presentation by: Burt Vialpando, Executive Analytics Architect, Hybrid Cloud Black Belt Team

Contributions by: Hemanth Manda, Director of Platform Offerings, IBM Data and AI
Vikram S Khatri, Executive Architect, Cloud Pak for Data
Travis Jeanneret, Senior Certified Architect

October 1, 2019

© 2019 IBM Corporation

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

<ul style="list-style-type: none">• Introduction and Setup	<ul style="list-style-type: none">• Lab 01
<ul style="list-style-type: none">• Executive Demo	<ul style="list-style-type: none">• Lab 02
<ul style="list-style-type: none">• Collect Part 1 – Connect• Organize• Collect Part 2 – Virtualize	<ul style="list-style-type: none">• Lab 03• Lab 04• Lab 05
<ul style="list-style-type: none">• Analyze Part 1 – Dashboards (optional)• Analyze Part 2 – Model Creation	<ul style="list-style-type: none">• Lab 06• Lab 07
<ul style="list-style-type: none">• Deploy and Infuse• Wrap-up	<ul style="list-style-type: none">• Lab 08• Lab 09

IBM Analytics Modernization Workshop

The workshop image desktop





Introduction and Setup

Lab 01 – Introduction and Setup

Technology by itself is not the business disruptor. Failing to be customer centric is the biggest business threat.

Netflix did not kill Blockbuster...

ridiculous late fees and rewind fees did.

Uber did not kill the taxi business...

limited access and fare control did.

Apple did not kill the music industry...

being forced to buy full length albums did.

Airbnb isn't killing the hotel industry...

limited availability and pricing options are.

**The right technology for the right job
is an enabler of business disruption**

Why Cloud Pak for Data?

The business case



Digital transformation, cloud & AI is disrupting every industry

- *75% of large enterprises will have digital transformation at the center of corporate strategy within the next 2 years*
- *>85% Of enterprise IT organization will commit to multi-cloud architectures*
- *>59% Of large enterprises see improved application quality & reduced defects using containers*



Increasing data volumes & diversity; Growing regulations; Outdated monolithic systems

- *81% of businesses are having issues preparing data required for AI*
- *Enterprise customers love public cloud, but are not yet ready to fully embrace it*

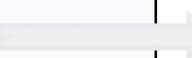


Cloud Pak for Data!

- *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!*

The need for “Borderless Data”

There are multiple borders that prevent Data Scientists from getting access to data. The borders around data come from access control, IT processes and un-curated data. These borders inhibit businesses from getting access to the data needed to drive meaningful insight.

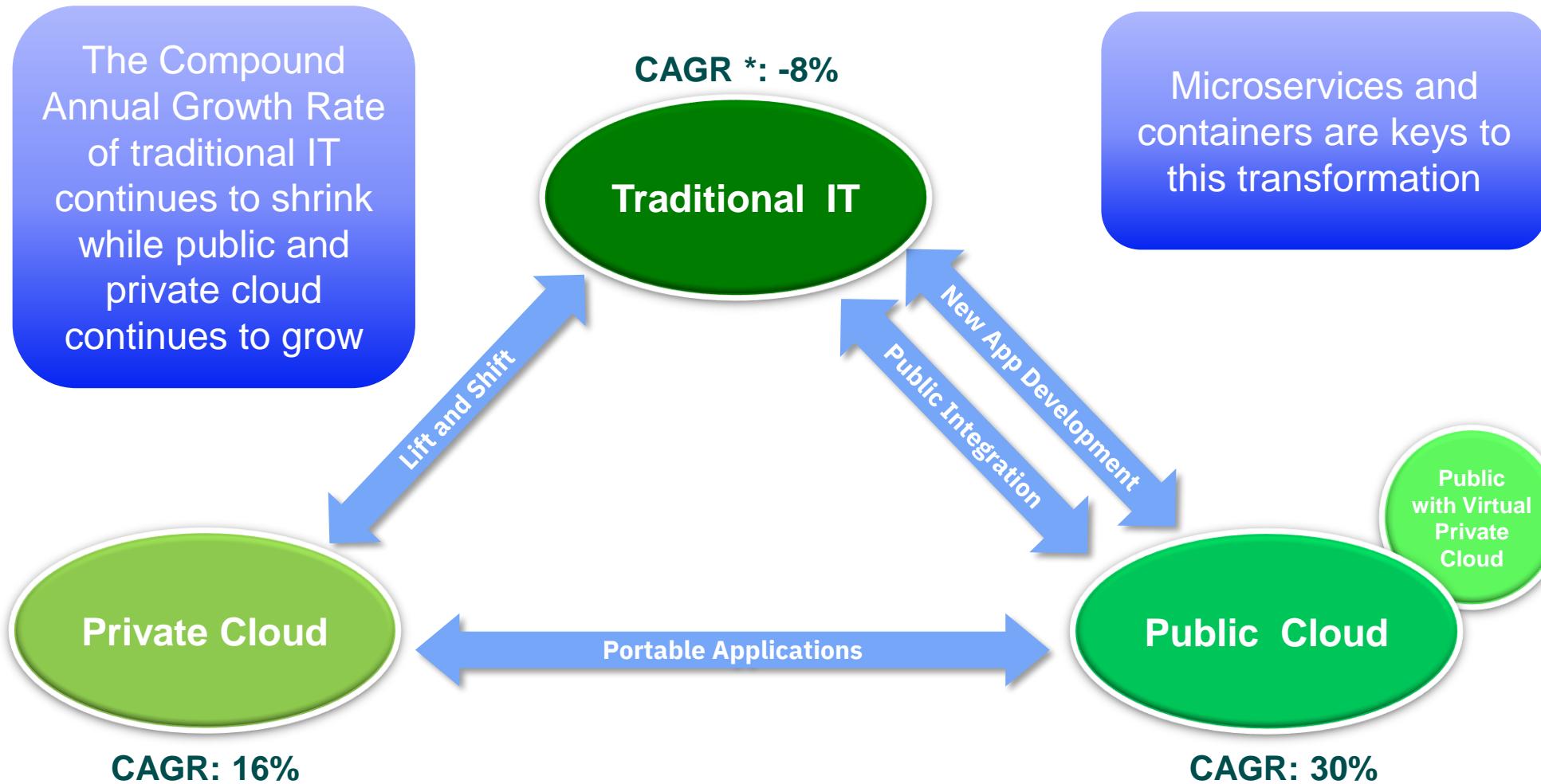


Harvest “Dark Data”

Gartner defines dark data as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing).

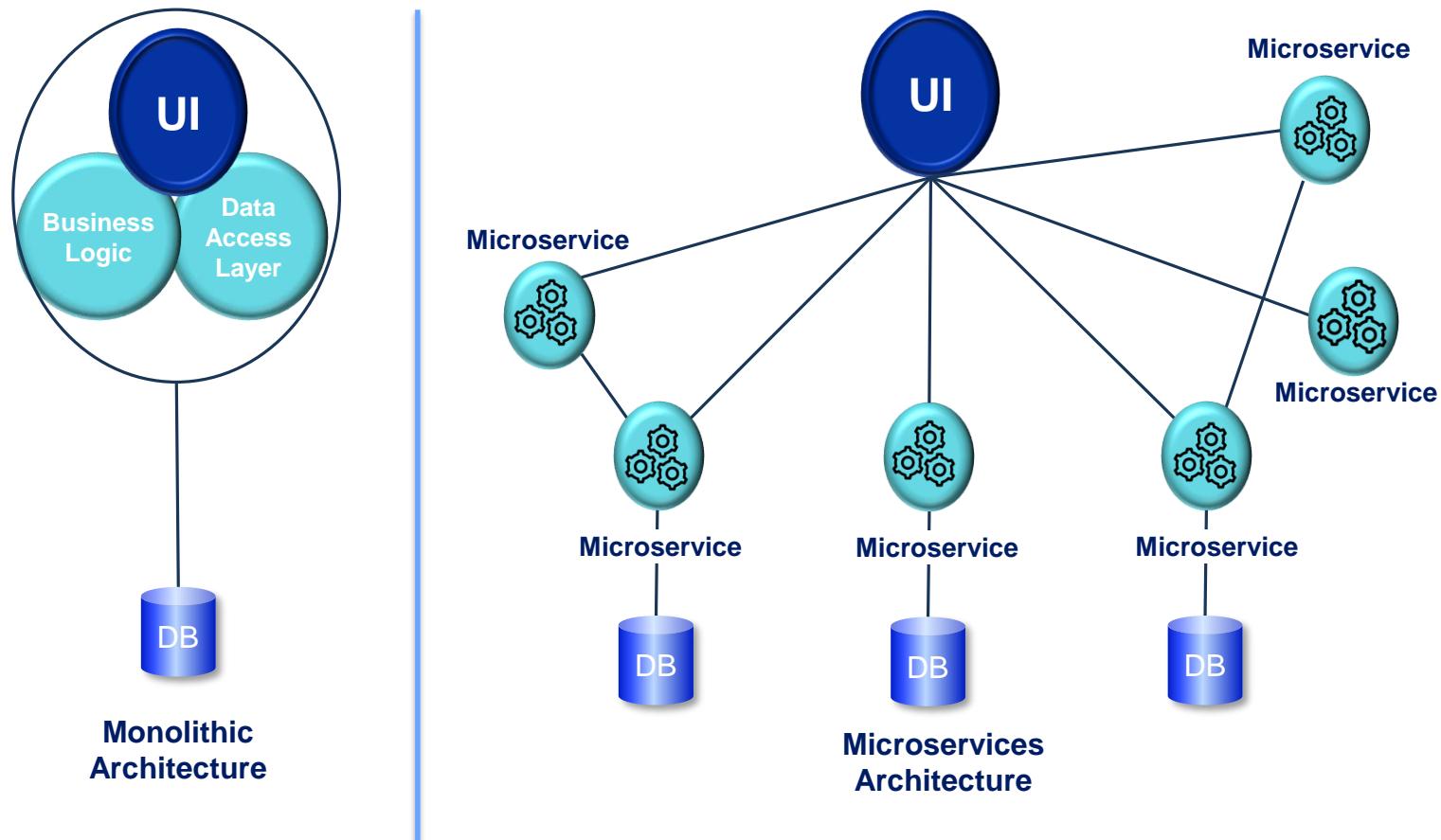
Multi-cloud is being driven by cloud native architectures

Microservices and containers are changing IT



Microservices – the first key to cloud native applications

Making development & deployment more efficient



Microservices benefits *

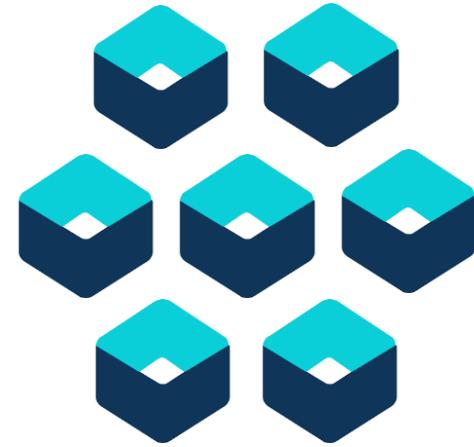
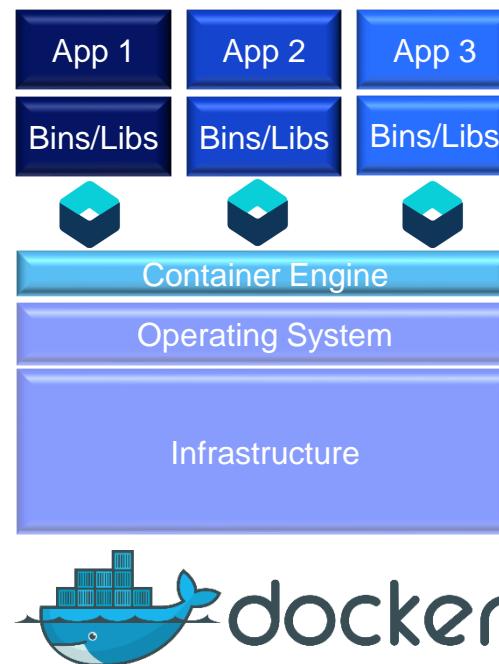
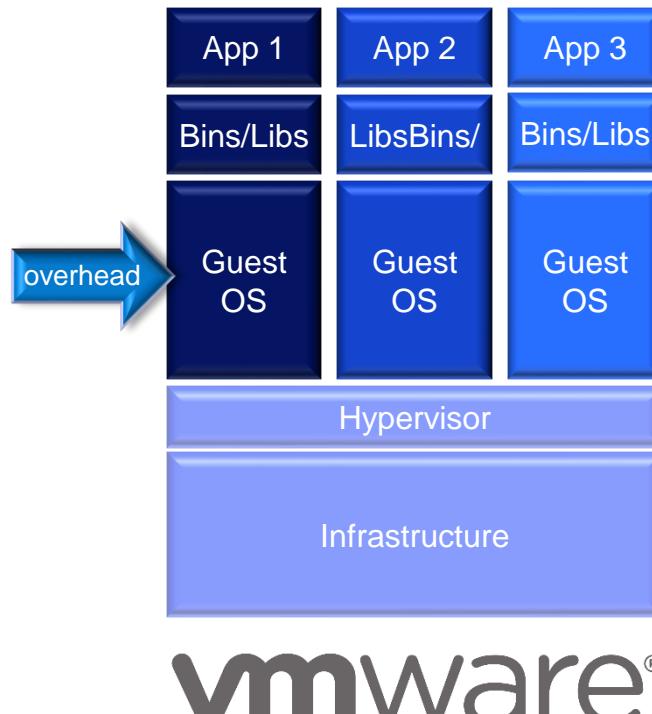
- **Improved fault isolation:**
Larger applications can remain largely unaffected by the failure of a single module
- **Technological flexibility:**
Try out a new technology stack on an individual service and roll it back if required
- **Easier development:**
A new developer can more easily understand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

bv

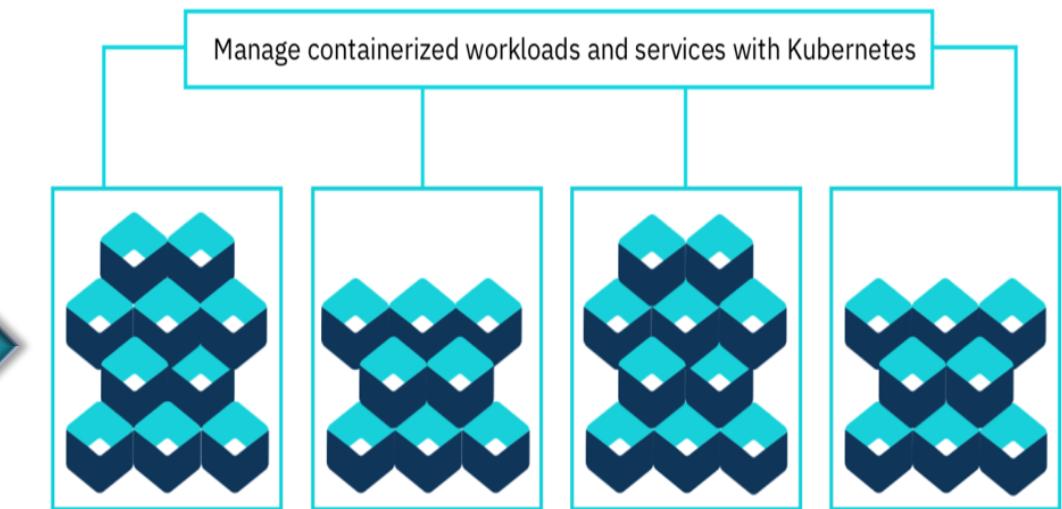
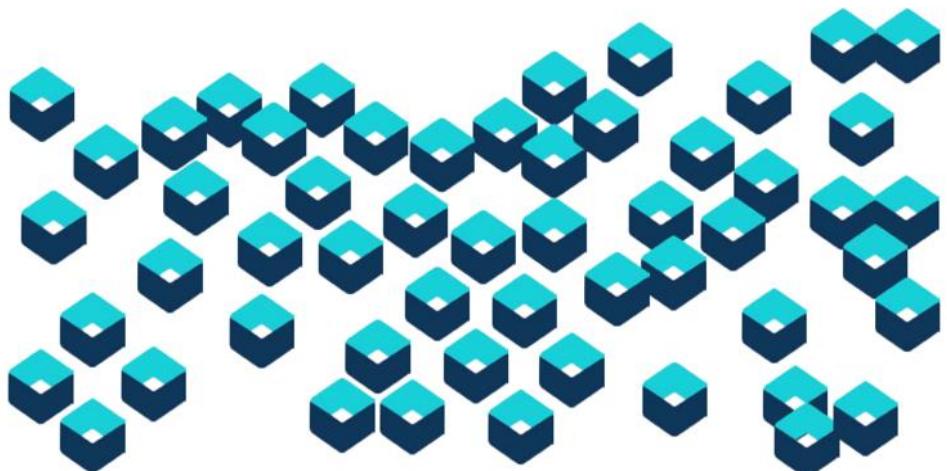
- ✓ Containers virtual software in the way that virtual machines have virtualized hardware

Container automation and orchestration is essential

Enter: Kubernetes 

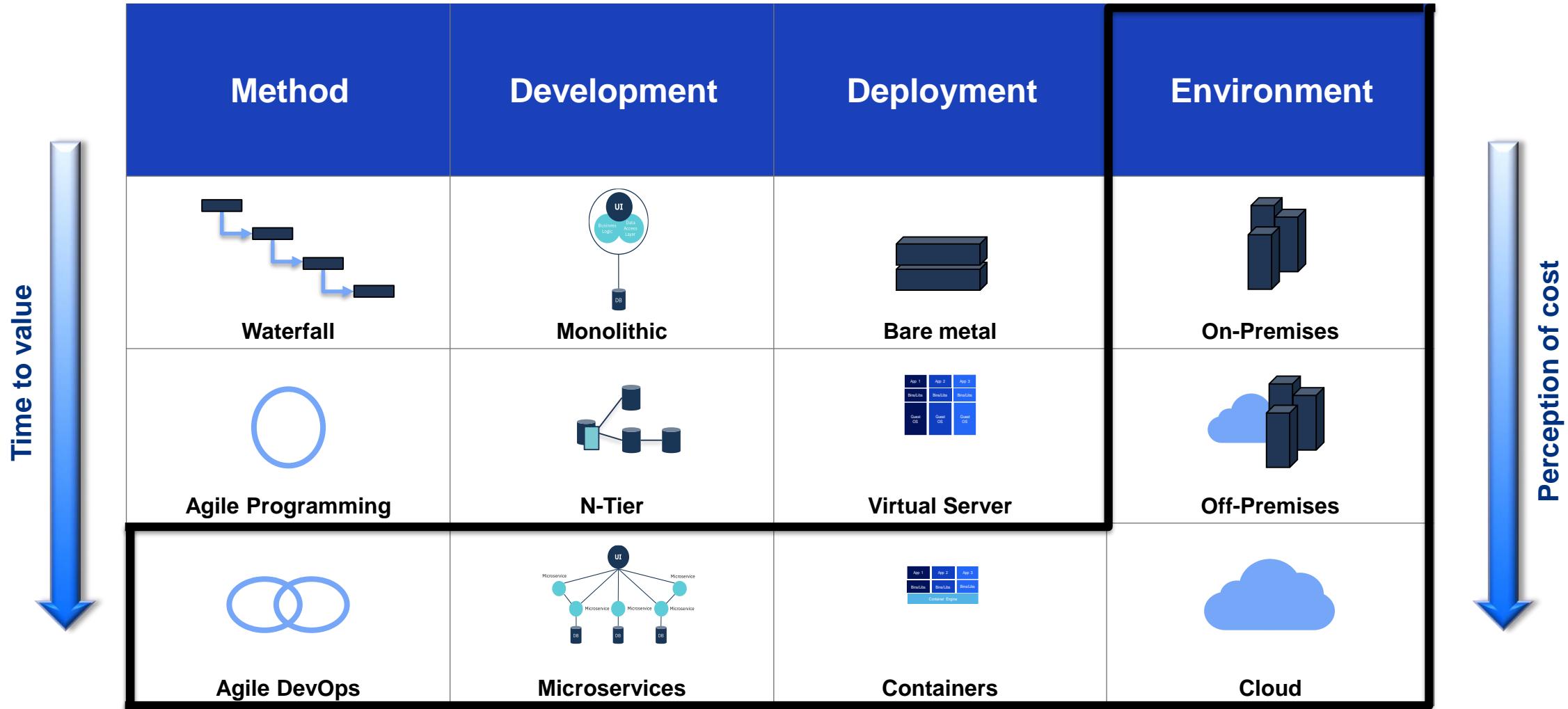
**Containers are revolutionizing IT
But they require orchestration**

Kubernetes - κυβερνήτης
Means “helmsman” or “pilot”



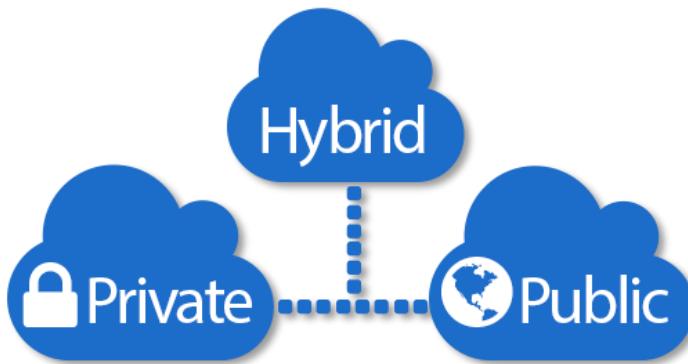
Private Clouds address the new IT reality

Created by digital transformation



Public Cloud + Private Cloud = Hybrid Cloud *

Different cloud options

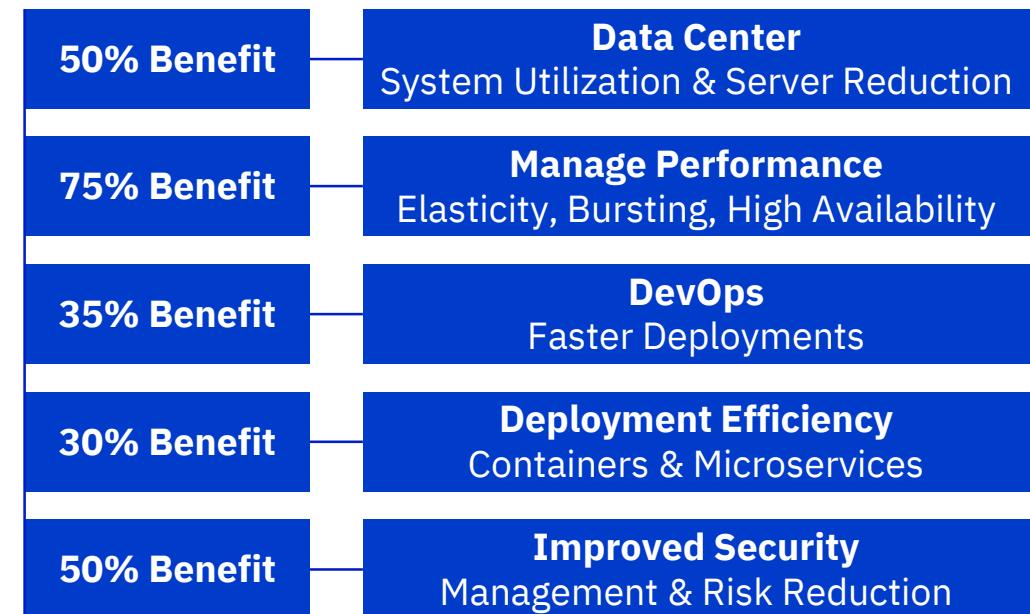
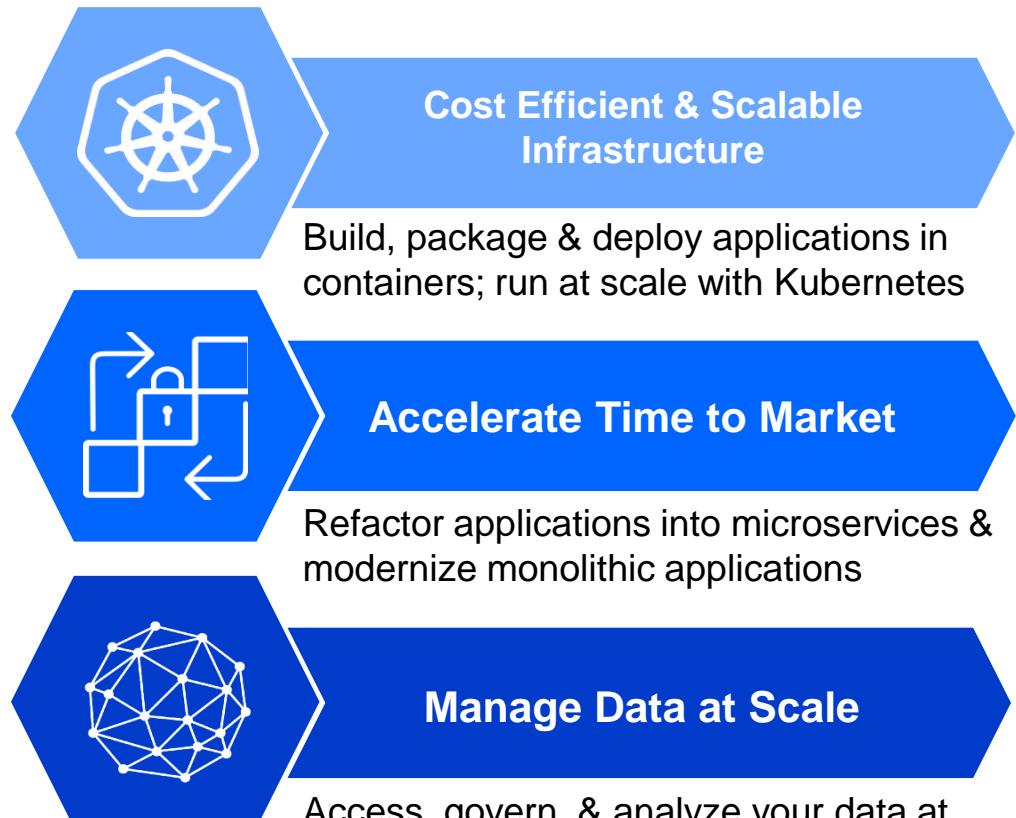


	Public Cloud	On-Premises Private Cloud	Hosted Private Cloud	Hybrid Cloud
Hardware Deployment and Management	Vendor	Customer	Vendor	Shared between vendor and customer
Hardware Sharing Model	Shared	Dedicated	Dedicated	Partially shared and partially dedicated
Scalability	High	Medium	High	High
Low Cost	Yes	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

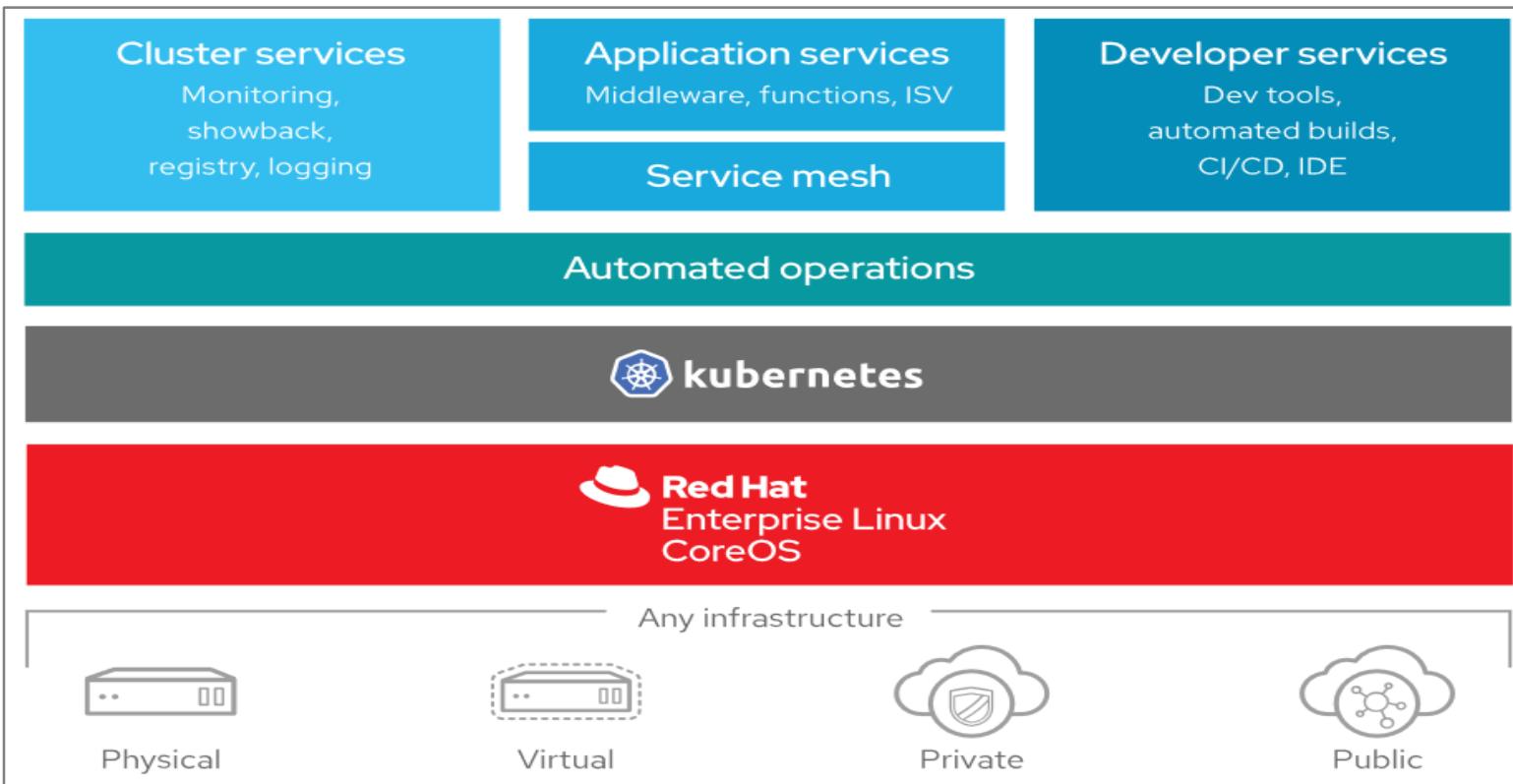


3-Year \$5.4 Million Cost Savings; 255% ROI

Business Value Assessment Customer Output:
Standard On-Premises vs IBM Cloud Private

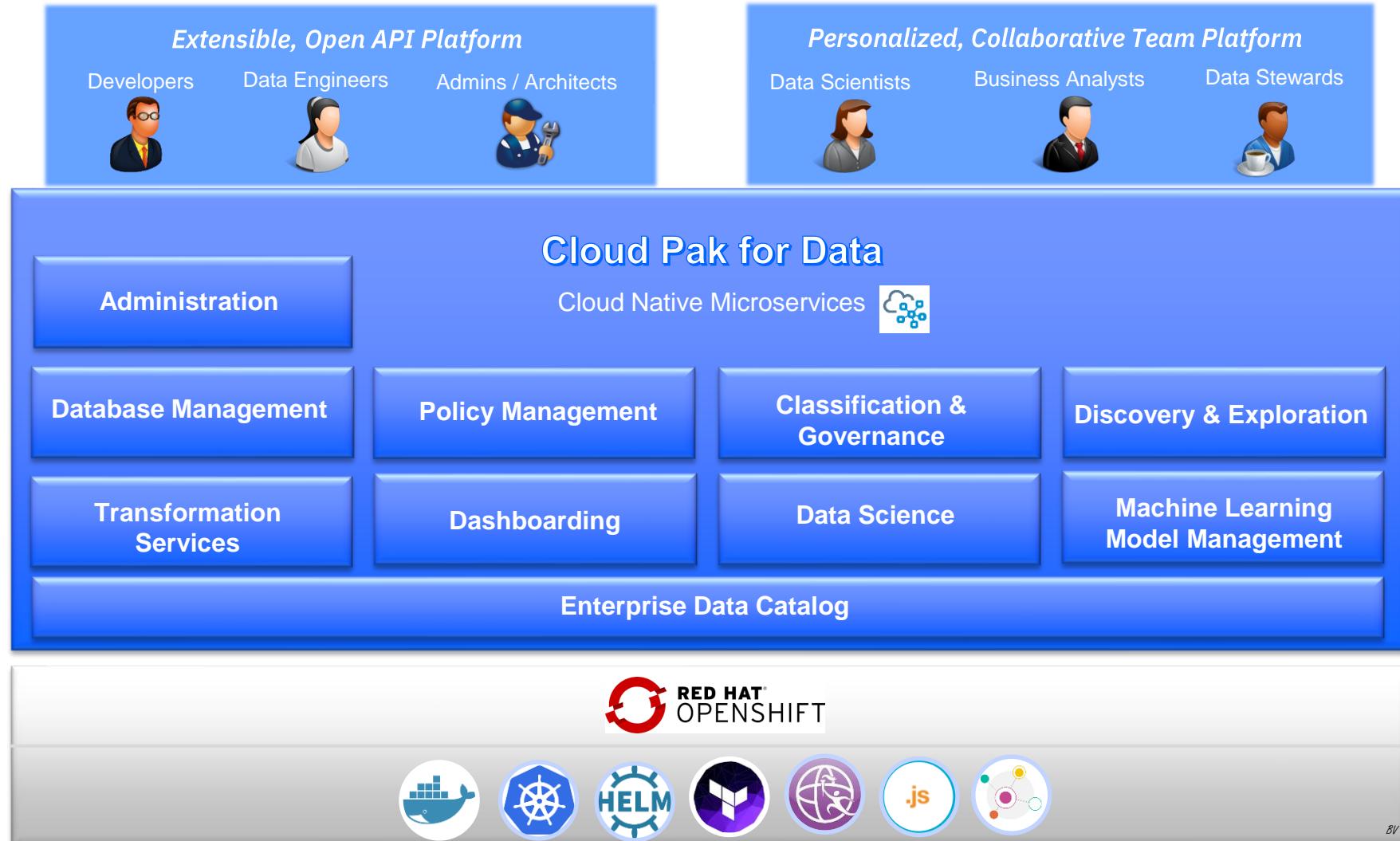
Private Cloud Platform Market Leader

OpenShift



Cloud Pak for Data (CPD)

Analytics Modernization platform built to run on Private Cloud



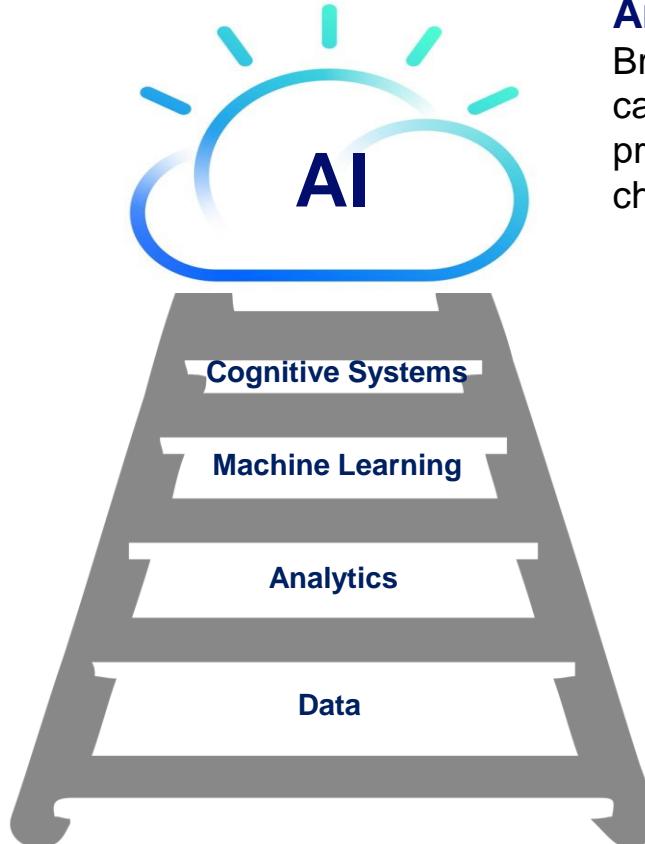
Cloud Pak for Data (CPD)

High level component overview

	Component Capability	Description
 Collect	Database Management	1) In-cluster databases, 2) Native connection databases and data source repositories, 3) Data Virtualization
	Policy Management	Business rules and validation of data assets, compliance, define quality standards in the catalog
	Classification & Governance	Auto classification of data assets in the catalog, assignment of terms to assets, tagging with ML & fuzzy/pattern based classifications for governance
	Discovery & Exploration	Database schema and table extractions, metadata syncs to feed into the Enterprise Catalog
	Transformation Services	Transformation jobs, working with data source definitions, management of ETL flows
	Dashboarding	Aggregate & display metrics & key performance indicators (KPIs), enabling them to be examined at a glance before further exploration
	Data Science	Access to open source tooling, frameworks, and IBM value-add
	Machine Learning Model Management	Deployment of machine learning models with elasticity and load balancing, monitoring, and management of models in production

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 General Manager, Watson and AI

Cloud Pak for Data (CPD)

Make your data ready for AI



1 Strong Foundation – Built on “Cloud native architecture”

2 The updated AI Ladder



Infuse – Deploy trusted AI-driven business processes



Analyze – Scale insights with ML everywhere



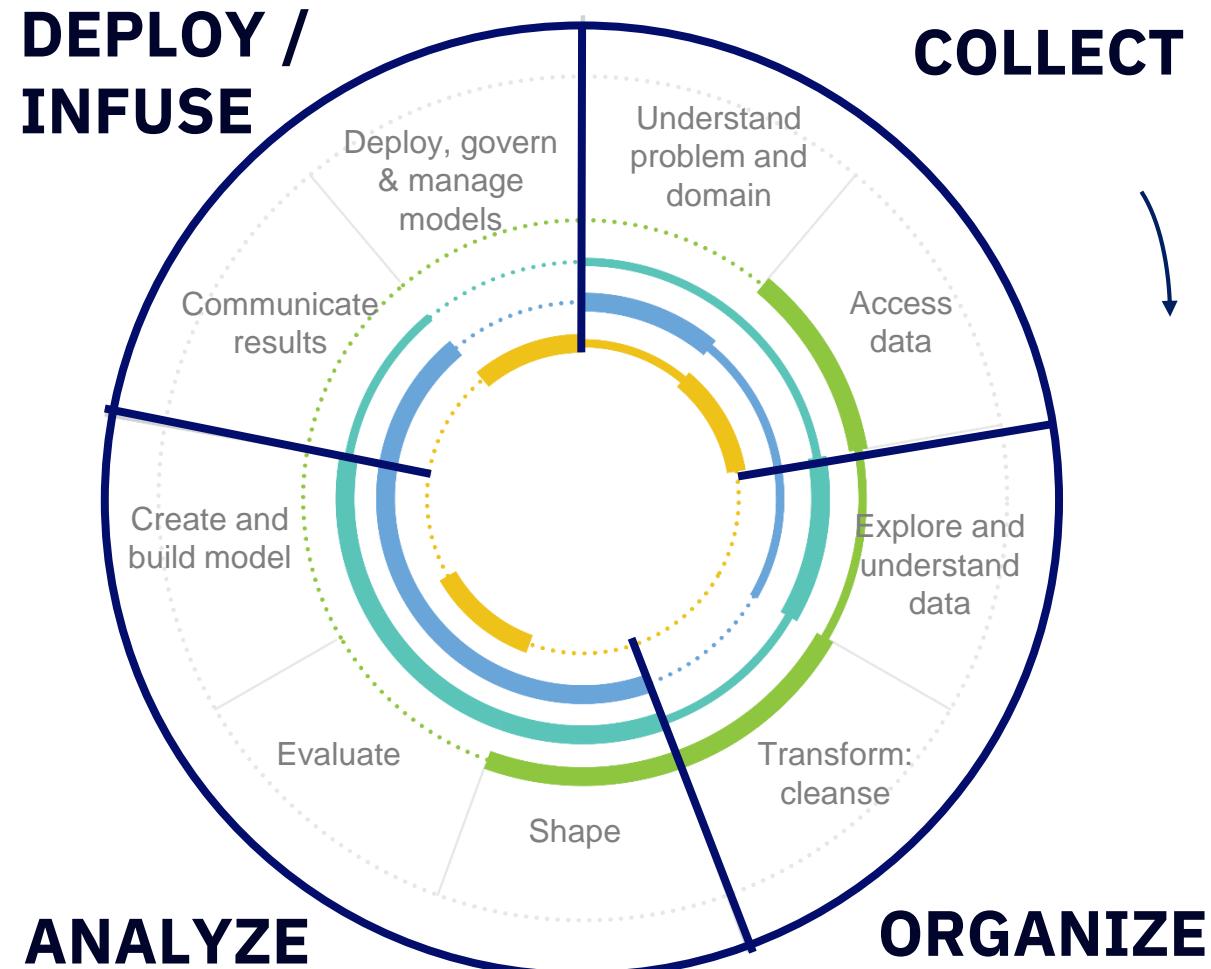
Organize – Create a trusted analytics foundation



Collect – Make data simple & accessible

Cloud Pak for Data (CPD)

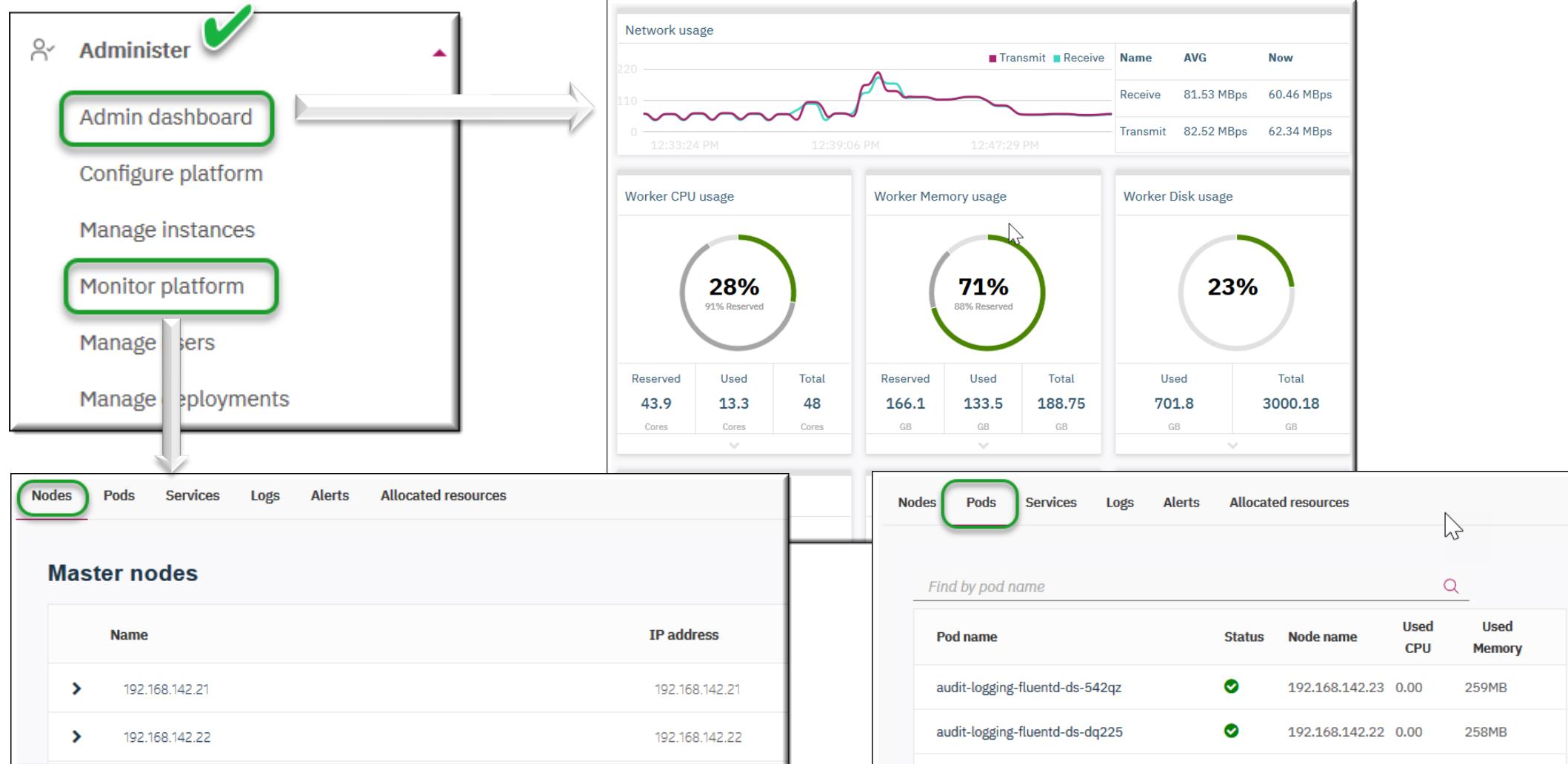
Increases workforce productivity across the analytics lifecycle



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

Administer and Monitor the platform



CPD

Manage environments and instances

My Instances

Environments **Provisioned instances** Jobs

My Images

Search by runtime name

Runtimes 1

Name	Runtime Type	User	Project	Job Name (Run Id)	Date
Jupyter with Python 3.6, Spark 2.3.2	Environment	admin	TradingCustomerChurn	—	24 Mar 10:35

Environments **Provisioned instances** Jobs



Name	Type	Provisioned by
MongoDB-Activity1	mongodb	user999
MongoDB-Activity2	mongodb	user999
Db2WarehouseSMP	db2whsmp	user999
Data Virtualization	dv	user999

Introduction and Setup

Lab 01

• Introduction and Setup	• Lab 01
• Executive Demo	• Lab 02
• Collect Part 1 – Connect	• Lab 03
• Organize	• Lab 04
• Collect Part 2 – Virtualize	• Lab 05
• Analyze Part 1 – Dashboards (optional)	• Lab 06
• Analyze Part 2 – Model Creation	• Lab 07
• Deploy and Infuse	• Lab 08
• Wrap-up	• Lab 09



Executive Demo

Lab 02 – Executive Demo

Boatswain Trading Company: Challenges



Customer retention problem leading to declining revenue

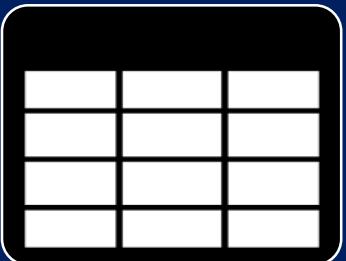
Underperforming rules-based system to identify separation (churn) risk

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

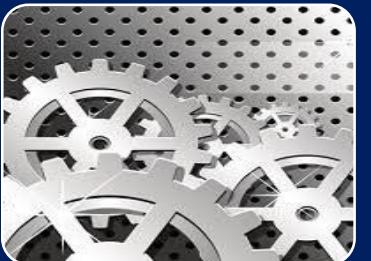
Low overall predictive accuracy. We are both missing identifying customers who ultimately separate and incorrectly assigning high risk to customers who ultimately stay.

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

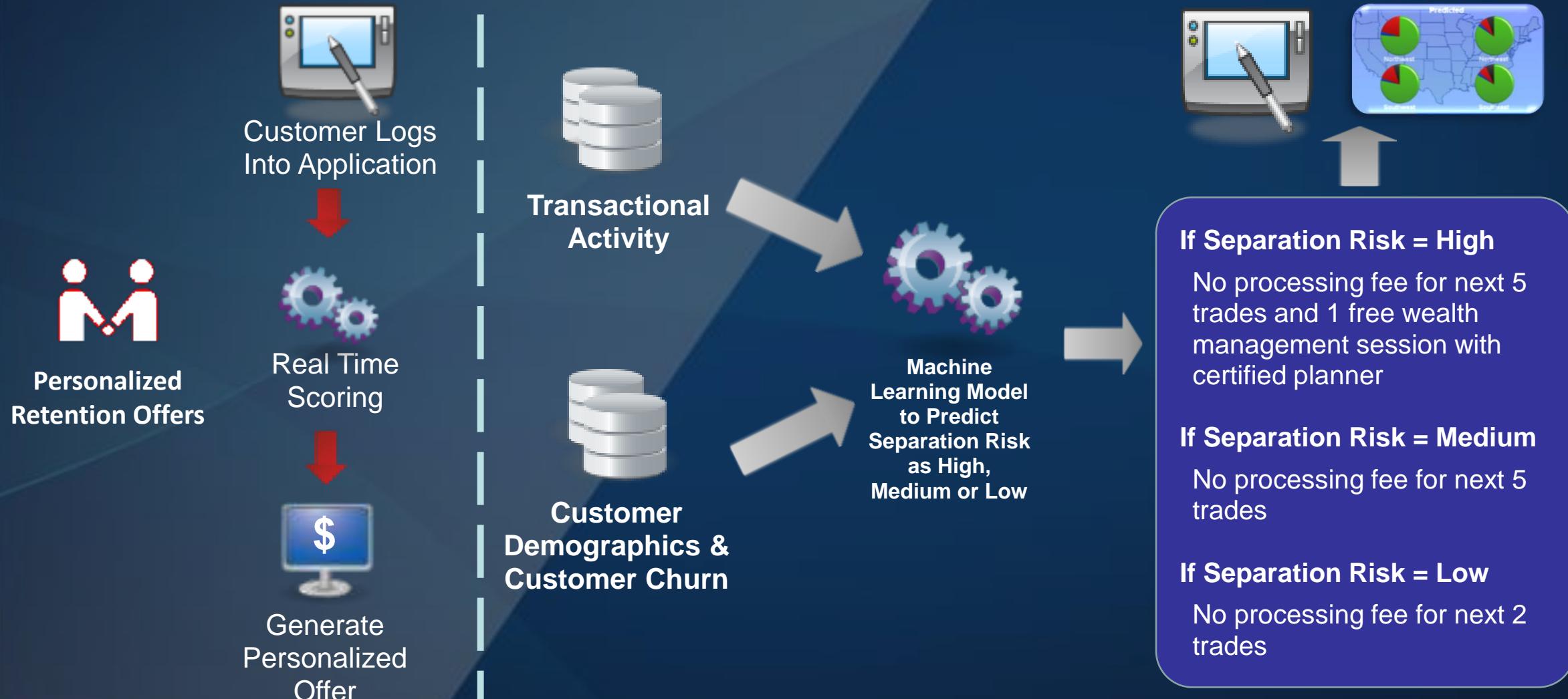
Develop predictive models for separation risk that automatically discover and incorporate all the patterns in the data including interactions and contingencies.



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 with Integrated AI



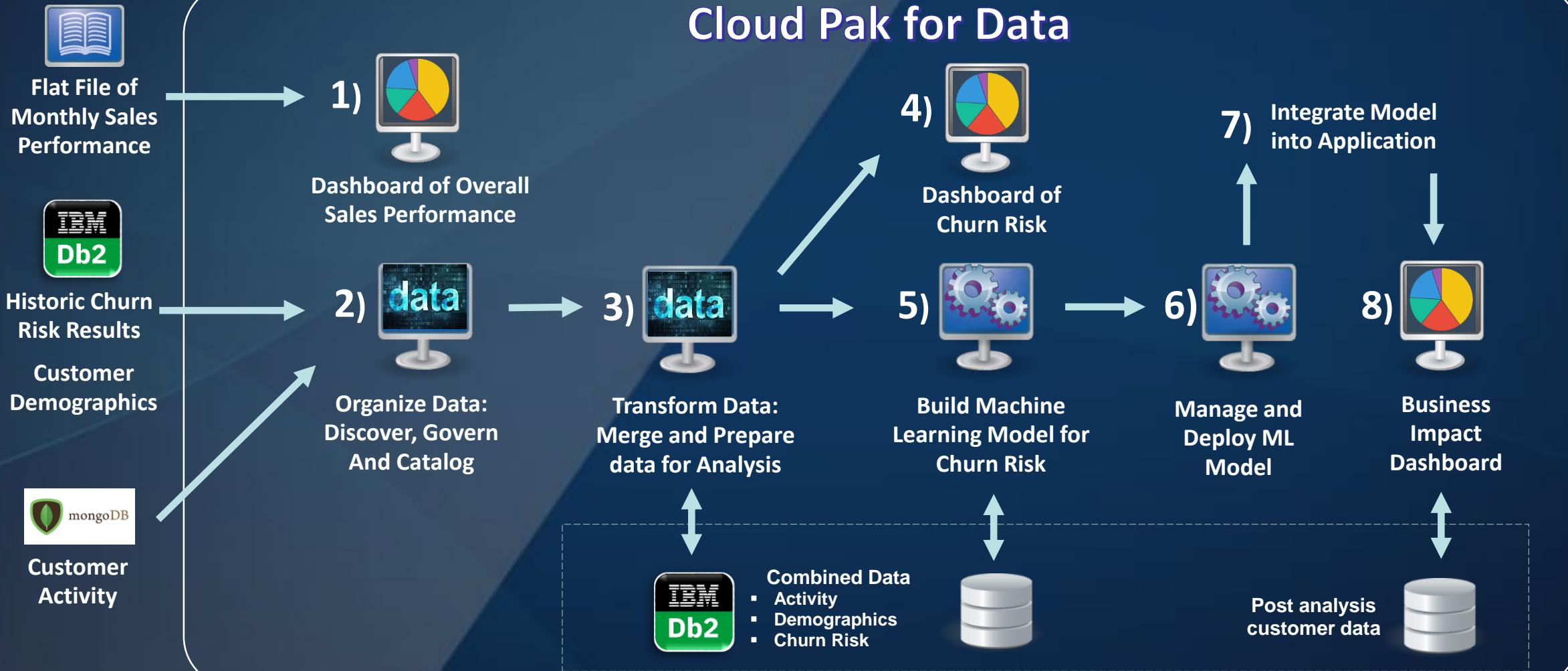


Stock Trader Application

Stock Trader Application
with Infused ML



Cloud Pak for Data



COLLECT

ORGANIZE

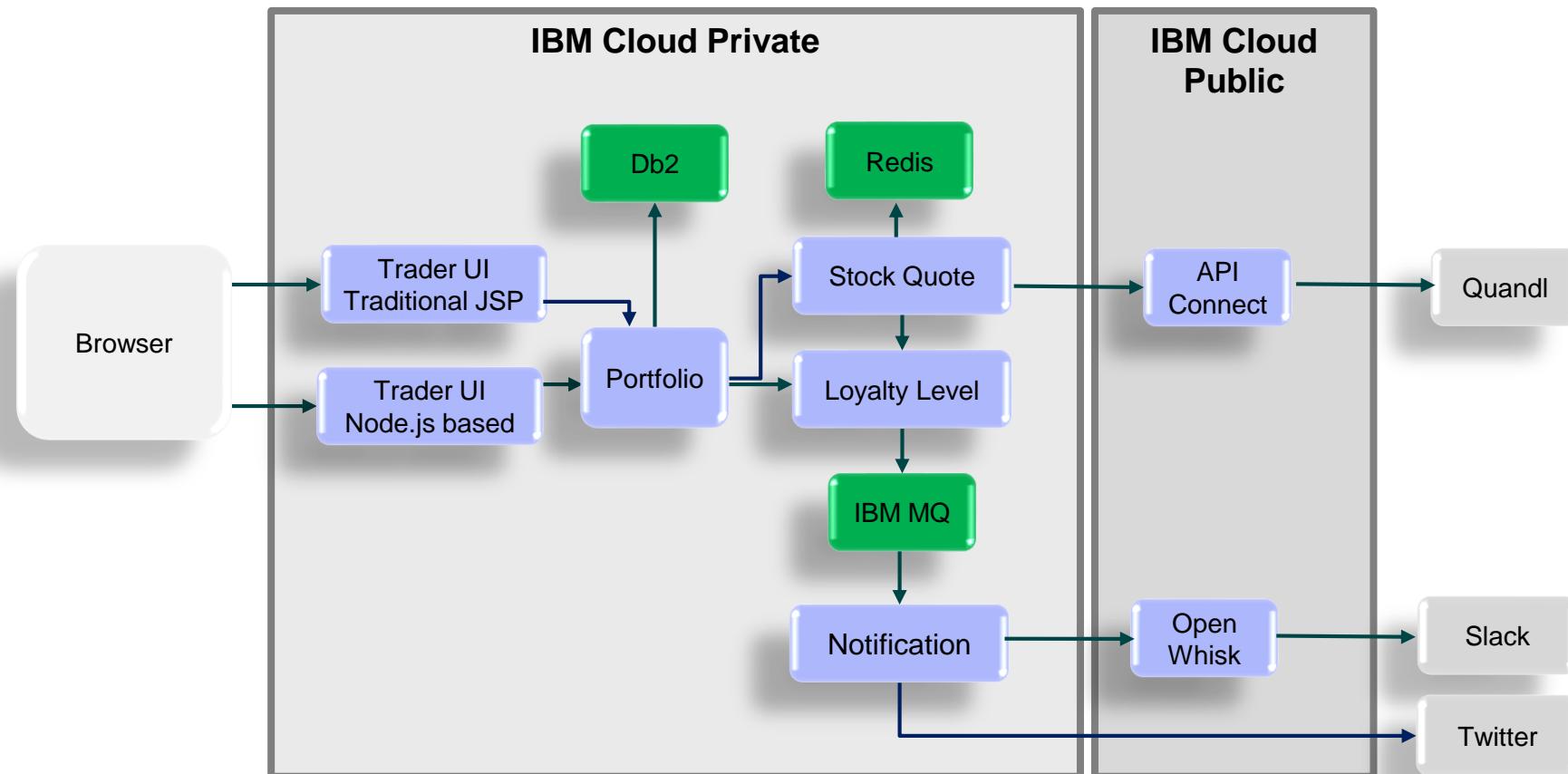
ANALYZE

DEPLOY / INFUSE

Stock Trader – Before Microservices Application

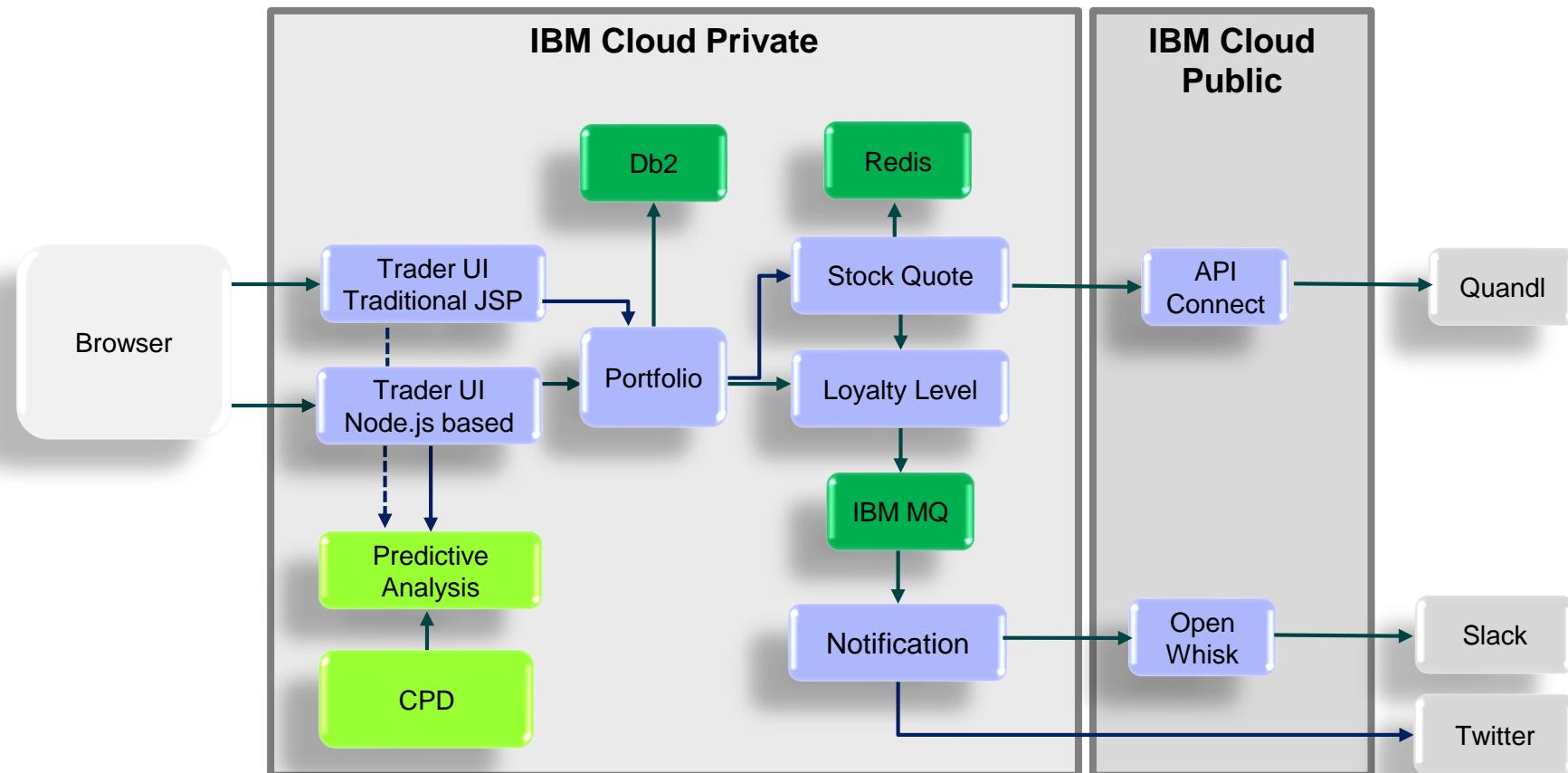
Stock Trader

as a modern microservice application



Stock Trader – After: Infused with Machine Learning Microservices Application

Stock Trader (enhanced)
as a modern microservice application



Stock Trader – After Monetizing the ML model

IBM TRADER

Home Summary Add Portfolio Predictive Analysis Change User

Summary

Welcome to IBM Trader powered by ICP for Data

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

Owner	Total	Loyalty Level
TechStocks	\$115,670	Gold

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

no processing fee for next 5 trades

Advertisement

IBM Cloud Private for Data

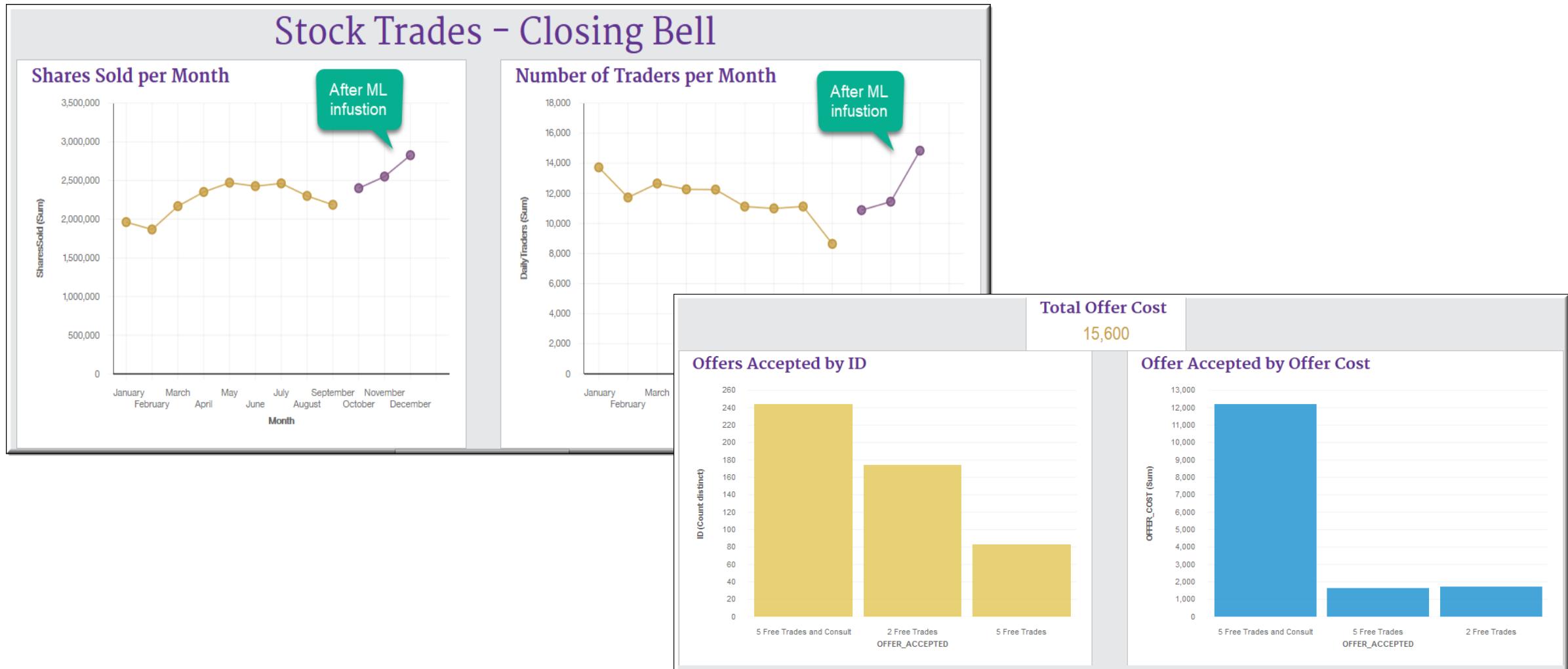
- Cloud agile
- Lightning fast
- AI-ready

No assembly required

Submit Change User

Stock Trader Closing Bell – Dashboard

Stock Trader – After deploying the CPD developed ML model



Executive Demo

Lab 02

<ul style="list-style-type: none">• Introduction and Setup• Executive Demo• Collect Part 1 – Connect• Organize• Collect Part 2 – Virtualize• Analyze Part 1 – Dashboards (optional)• Analyze Part 2 – Model Creation• Deploy and Infuse• Wrap-up	<ul style="list-style-type: none">• Lab 01• Lab 02• Lab 03• Lab 04• Lab 05• Lab 06• Lab 07• Lab 08• Lab 09
--	--



Collect and Organize

Lab 03 – Collect Part 1 – Connect

Lab 04 – Organize

Lab 05 – Collect Part 2 – Virtualize

CPD Collect

1. Provision in-cluster databases

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

 CockroachDB Partner Premium Delivers the reliability of a transactional RDBMS, the scale of NoSQL and the ease, durability and ubiquity of the cloud.	 Db2 Event Store IBM Premium An in-memory database that is designed to handle massive amounts of structured data that is stored in Apache Parquet format. Because it is optimized for event-driven data processing and analysis, this high-speed data store can capture, analyze, and store more than 250 billion events per day.	 Db2 Advanced Enterprise Server Edition IBM Premium Provides advanced data management and analytics capabilities for transactional workloads. It has no processor, memory, or database size limits, which makes it ideal for any size of workload.	 Db2 Warehouse MPP IBM Client managed analytics warehouse, optimized for fast and flexible deployment with automated scaling to meet agile analytic workloads with in-memory BLU processing technology and in-database analytics, plus scalability and performance through the MPP architecture.
 Db2 Warehouse SMP IBM Enabled Client managed analytics warehouse, optimized for fast and flexible deployment with automated scaling to meet agile analytic workloads with in-memory BLU processing technology and in-database analytics.	 IBM Db2 for z/OS IBM Delivers real-time insight by bringing analytics to your data and transactions. Chosen for its proven resiliency, reliability, security, and scalability, it also reduces latency, complexity, and cost while improving data quality and governance.	 MongoDB Enterprise Partner Enabled Premium An open source cross-platform document-oriented NoSQL database that eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas, which makes it easier and faster to integrate data in certain types of applications.	<p>V2.1.0 adds:</p> <ul style="list-style-type: none">✓ PostgreSQL✓ IBM Streams

CPD Collect

2. Connect to existing data sources (v2.1.0)

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

Data Source	Governance	Transformation (ETL)	Data virtualization	Analytics projects	Analytics dashboard	Data Source	Governance	Transformation (ETL)	Data virtualization	Analytics projects	Analytics dashboard
Amazon S3		✓				Hyperledger				✓	
Apache Derby			✓			Informix®			✓	✓	
BDFS		✓				JDBC		✓			
Big SQL			✓	✓	✓	Kafka		✓			
Cloudera Impala			✓			MariaDB			✓		
3rd party JDBC drivers			✓	✓		Microsoft SQL Server	✓	✓	✓	✓	✓
Db2® Event Store	✓	✓	✓			MongoDB	✓		✓		
Db2 Warehouse on Cloud	✓	✓	✓	✓	✓	MySQL Community Edition			✓		
Db2 on Cloud	✓	✓	✓	✓	✓	MySQL Enterprise Edition			✓		
Db2 (for LUW)	✓	✓	✓	✓	✓	Netezza®		✓	✓	✓	
Db2 for z/OS®	✓	✓	✓	✓		ODBC		✓			
Greenplum			✓			Oracle	✓	✓	✓	✓	✓
HDFS - Generic WebHDFS	✓					PostgreSQL			✓		
HDFS – HttpFS	✓					Sybase			✓		
HDFS – CDH						Teradata	✓	✓	✓		
HDFS – HDP						WebSphere® MQ		✓			
Hive JDBC			✓			Data File	Governance	Transformation (ETL)	Data virtualization	Analytics projects	Analytics dashboard
Hive HCAT – CDH				✓		CSV		✓	✓	✓	
Hive HCAT – HDP				✓		Microsoft Spreadsheets			✓		
Hive JDBC – CDH	✓	✓				Sequential file		✓			
Hive JDBC – HDP	✓	✓									

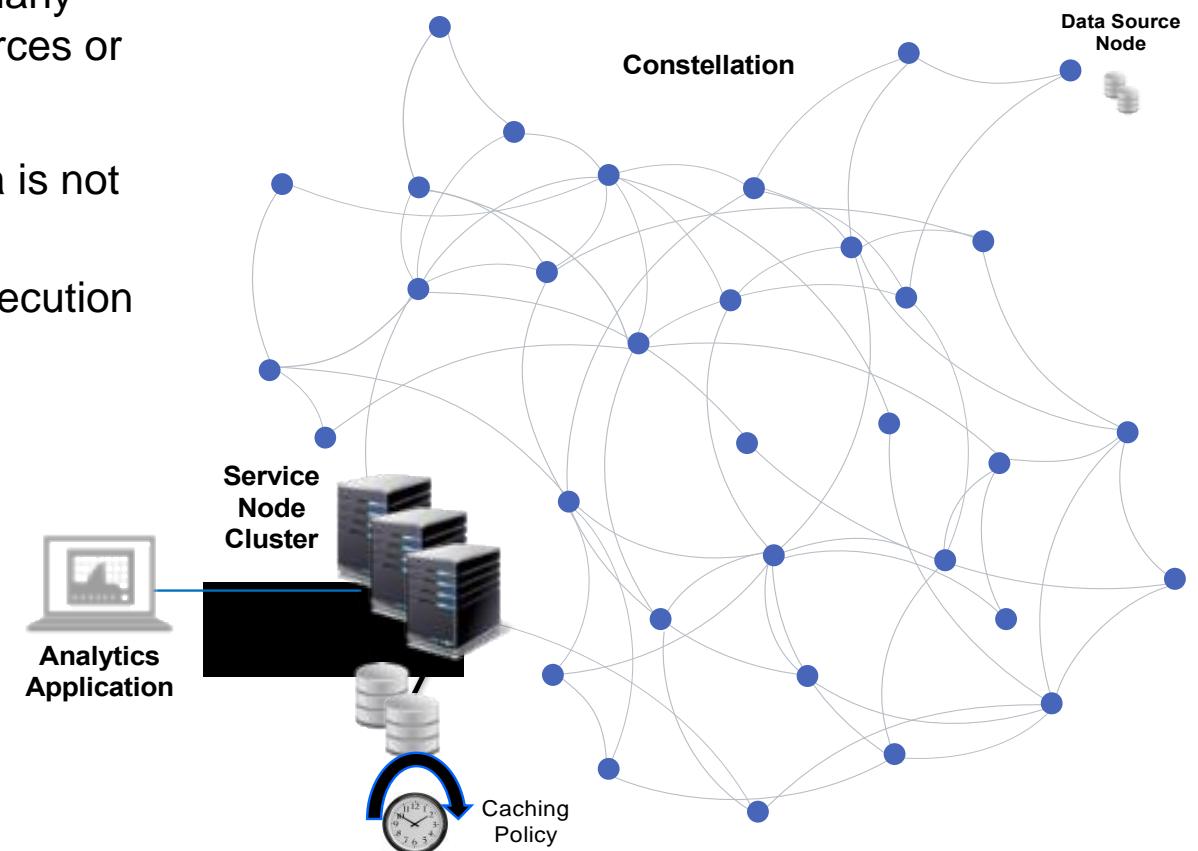
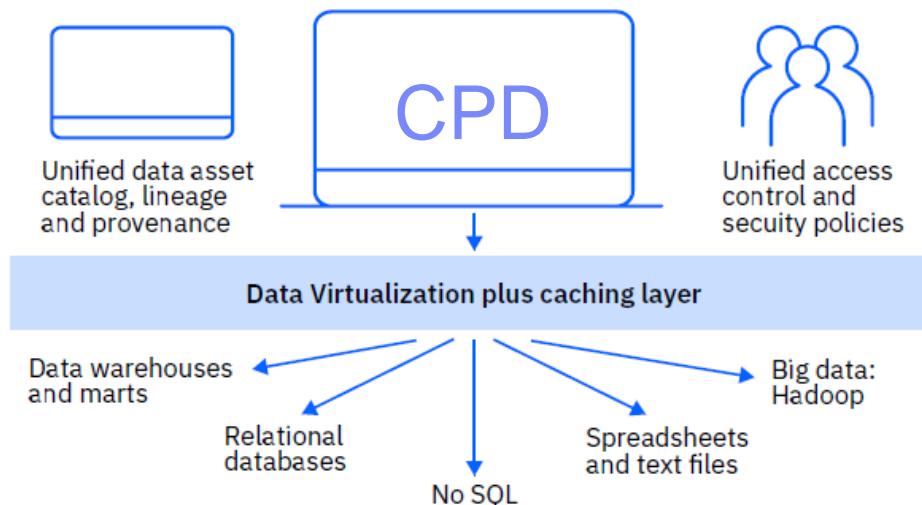
CPD Collect

3. Data Virtualization - Overview

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

Data Virtualization 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*

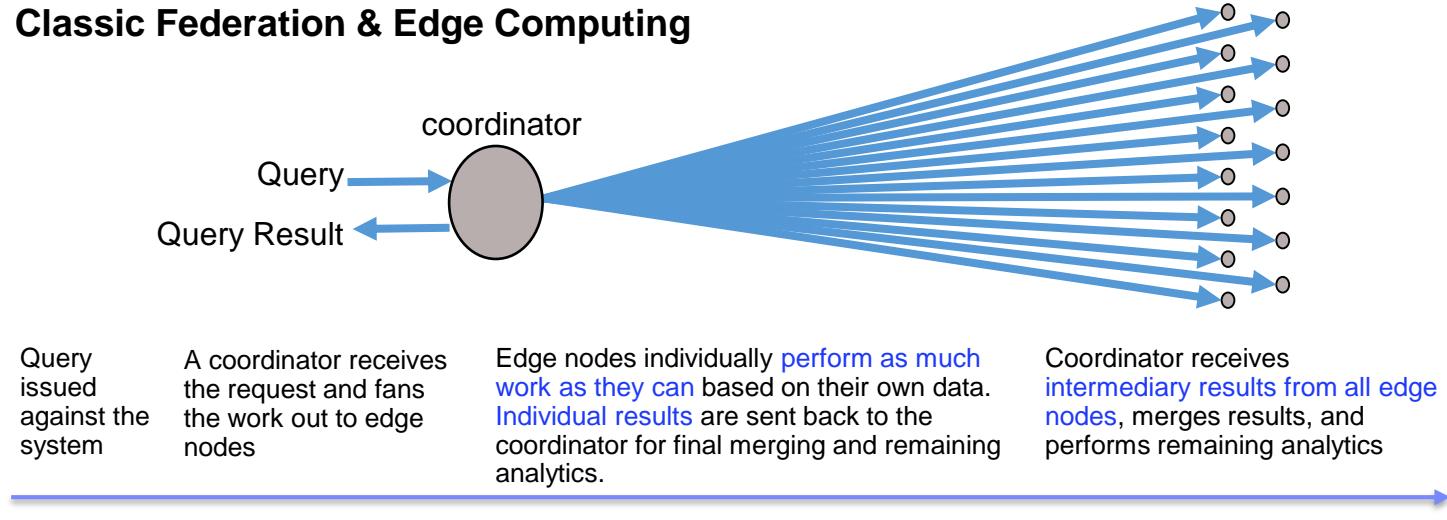
- Analytics are submitted against the data at the source (data is not copied)
- New algorithms are applied that support distributed SQL execution across heterogenous data sources



CPD Data Virtualization

Constellation “Computational Mesh” benefit

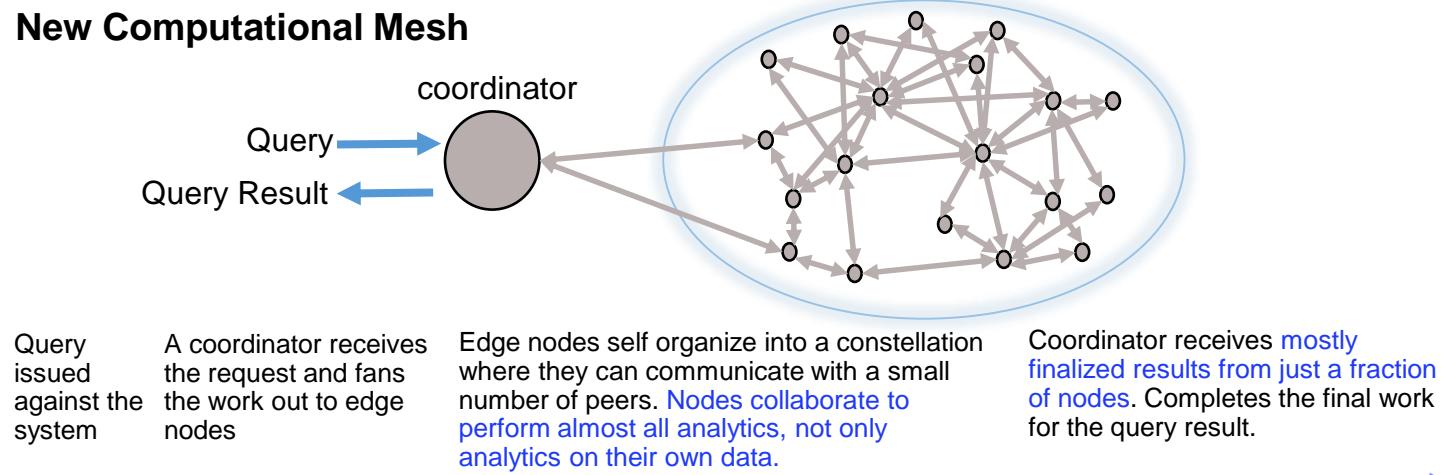
Classic Federation & Edge Computing



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

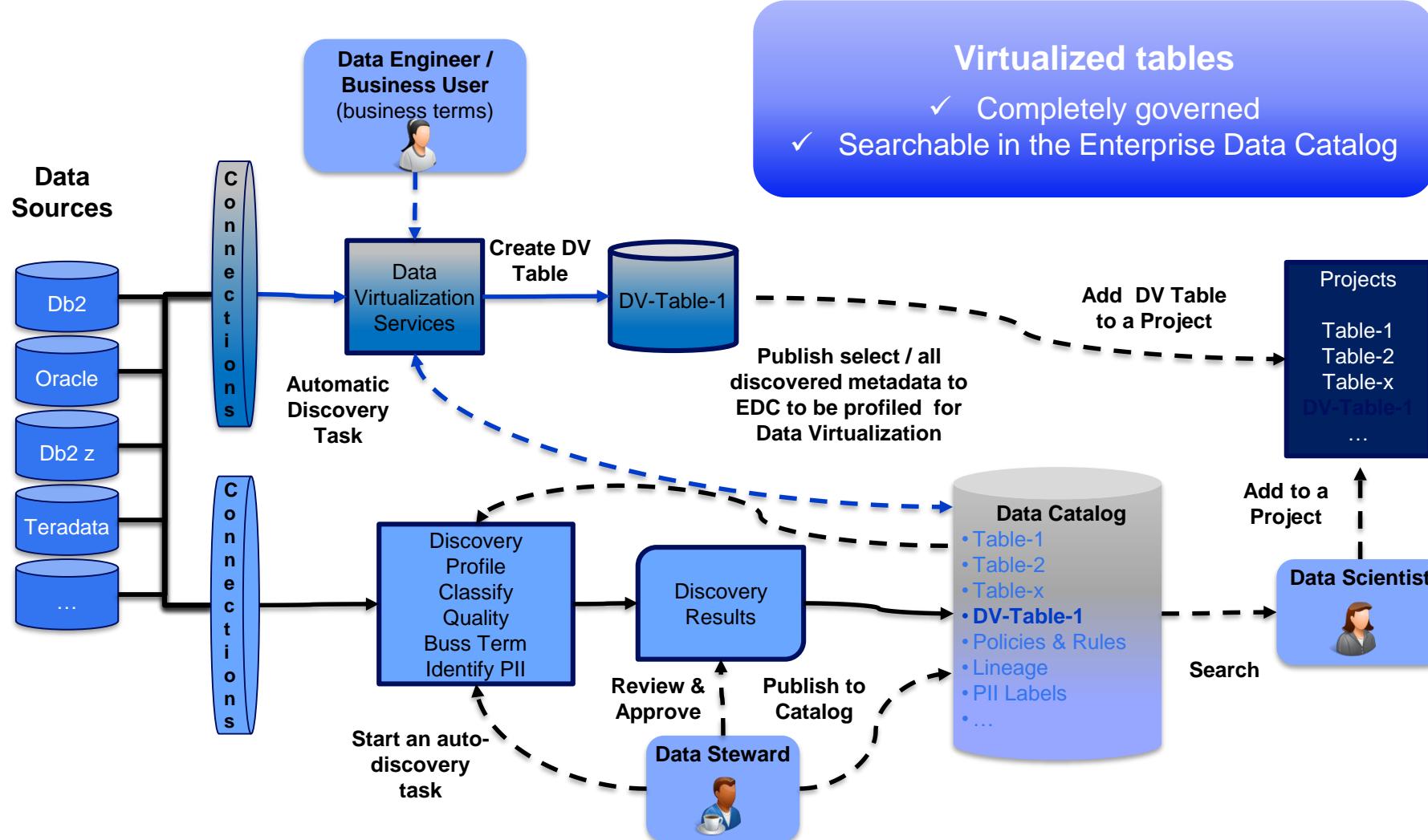
CPD Data Virtualization uses a new Computational Mesh approach which meets the performance demands of today's modern data access requirements

New Computational Mesh



CPD Data Virtualization

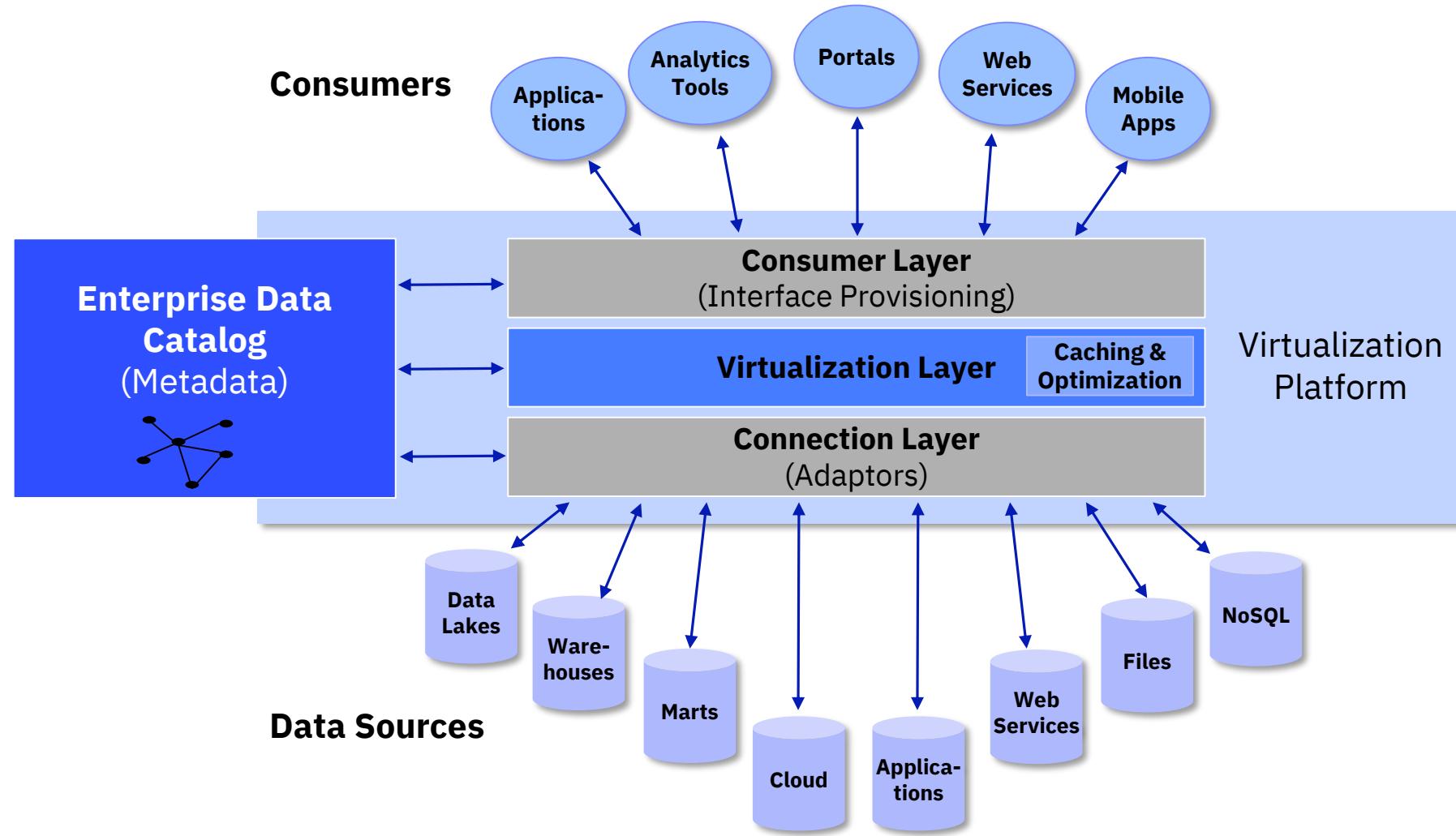
Data Governance is built in



- 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

Enterprise Data Catalog is 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:

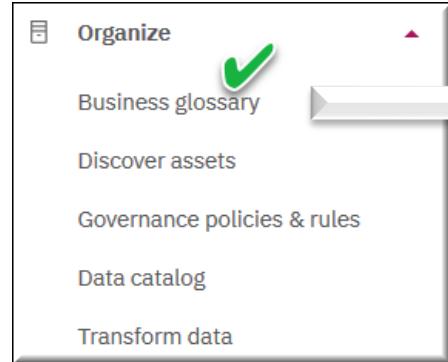
- Current state required for up-to-date analytics
- One time access to data, then throw it away
 - e.g. “How much cash is ‘on hand’ across our branches worldwide?” “What is our current ‘claims’ liability?”

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 Organize

Business Glossary: Create Categories and Terms



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

Industry accelerators

Credit Card Fraud IBM Banking Quickly detect credit card fraud to reduce financial losses and protect you and your customers. The Credit Card Fraud business glossary can jump-start your analysis.	Customer Churn Management IBM Cross-Industry Why are your customers leaving? Use the Customer Churn Management business glossary to jump-start your analysis.	Contact Center Optimization IBM Cross-Industry Need to improve the productivity of your customer contact center? The Contact Center Optimization business glossary can jump-start your analysis.	Customer 360 Degree View IBM Cross-Industry Need a complete picture of your customer base? The Customer 360 Degree View business glossary can jump-start your analysis.	Loan Default Analysis IBM Banking Identify potential credit risks in your loan portfolio. The Loan Default Analysis business glossary can jump-start your analysis.
---	---	--	---	---

- V2.1.0 adds:
- ✓ Customer Life Event Prediction
 - ✓ Customer Segmentation

CPD Organize

Define and enforce governance policies and rules

Rules Policies ✓

Search for Policies in the catalog

80 Policies available

Import Policies Create Policy

NAME	CREATED	MODIFIED	⋮
Information Protection	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Authoritative Sources provide the best information	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Policy Enforcement	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Information Supply Chain Integrity	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮

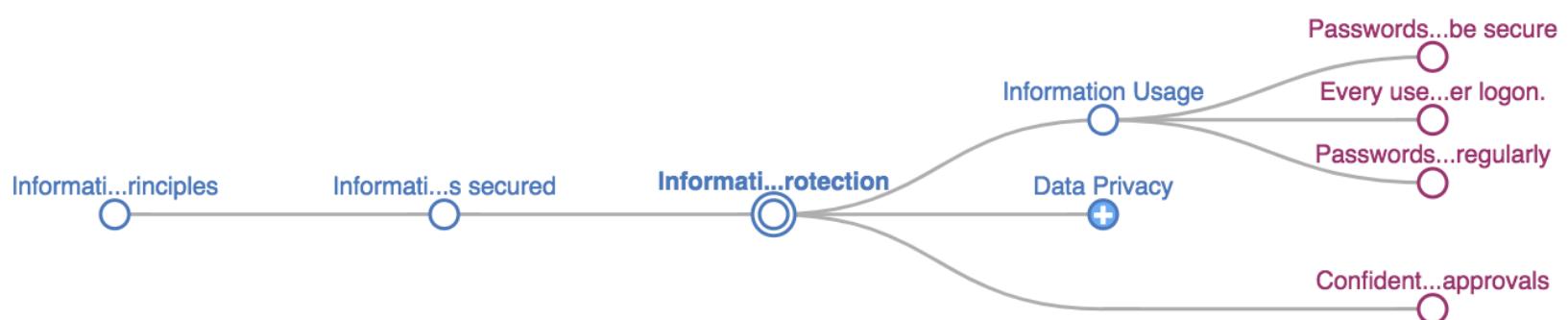
Rules Policies ✓

Search for Rules in the catalog

125 Rules available

Import Rules Create Rule

NAME	CREATED	MODIFIED	⋮
Mask an email address by using an auto generated method	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Mask a US Social Security Number by using a random method	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Mask a French INSEE Number by using a random method	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮
Computer addresses have a valid format	11 Sep 2018, 1:06 PM	11 Sep 2018, 1:06 PM	⋮



CPD Organize

Governance policy detail

IBM Cloud Private for Data

Protect Cardholder Data

Information Governance Policy Details

Search the catalog

Header

Header

Name

Protect Cardholder Data

Short Description

Protect stored cardholder data

Long Description

Protection methods such as encryption, truncation, masking, and hashing are critical components of cardholder data protection. If an intruder circumvents other security controls and gains access to encrypted data, without the proper cryptographic keys, the data is unreadable and unusable to that person. Other effective methods of protecting stored data should be considered as potential risk mitigation opportunities. For example, methods for minimizing risk include not storing cardholder data unless absolutely necessary, truncating cardholder data if full PAN is not needed, and not sending unprotected PANs using end-user messaging technologies, such as e-mail and instant messaging. Please refer to the PCI DSS and PA-DSS Glossary of Terms, Abbreviations, and Acronyms for definitions of strong cryptography and other PCI DSS terms.

Created by

admin admin

Created on

02 May 2018, 3:57:52 am

Modified by

admin admin

Modified on

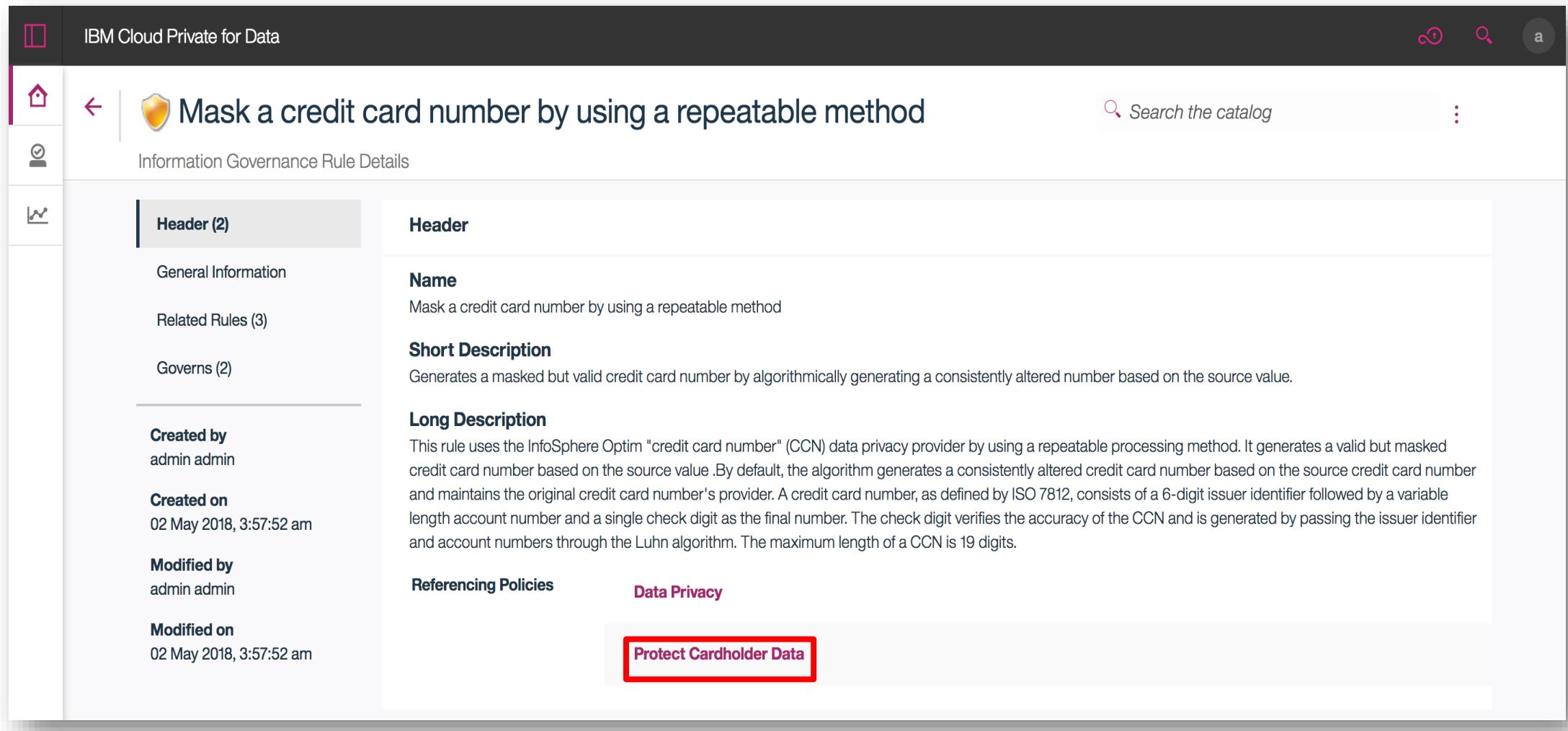
02 May 2018, 3:57:52 am

Parent Policy

[Information Governance Obligations](#) » [Industry Standards](#) » [Payment Card Information](#)

CPD Organize

Governance rule detail – masking example



The screenshot shows the 'Information Governance Rule Details' page for a rule named 'Mask a credit card number by using a repeatable method'. The left sidebar lists sections like General Information, Related Rules, and Governance. The main content area shows the rule's name, short description, long description, and referencing policies. A red box highlights the 'Protect Cardholder Data' button under Data Privacy.

IBM Cloud Private for Data

Mask a credit card number by using a repeatable method

Information Governance Rule Details

Header (2)

Header

Name

Mask a credit card number by using a repeatable method

Short Description

Generates a masked but valid credit card number by algorithmically generating a consistently altered number based on the source value.

Long Description

This rule uses the InfoSphere Optim "credit card number" (CCN) data privacy provider by using a repeatable processing method. It generates a valid but masked credit card number based on the source value. By default, the algorithm generates a consistently altered credit card number based on the source credit card number and maintains the original credit card number's provider. A credit card number, as defined by ISO 7812, consists of a 6-digit issuer identifier followed by a variable length account number and a single check digit as the final number. The check digit verifies the accuracy of the CCN and is generated by passing the issuer identifier and account numbers through the Luhn algorithm. The maximum length of a CCN is 19 digits.

Created by admin admin

Created on 02 May 2018, 3:57:52 am

Modified by admin admin

Modified on 02 May 2018, 3:57:52 am

Referencing Policies

Data Privacy

Protect Cardholder Data

CPD Organize

Auto-discover assets to catalog, publish meta data and track lineage

Data Discovery

Assets

Name	Quality	Data class
CUSTOMER_CHURN	91%	-
CHURNRISK	81%	City 82% ▾
ID	100%	Identifier 100% ▾
CUSTOMER_DEMOGRAPHICS	100%	-
GENDER	100%	Gender 100% ▾

Use machine learning based auto-discovery to:

1. Profile and classify data
2. Analyze data quality
3. Assign business terms to the data sources

You can perform this with data sampling to allow for self-service data access and a true 'shop for data' set of capabilities.

Explore Assets
In IBM Cloud Private for Data

ASSET TYPES

- Search asset types 🔍
- Glossary and Governance
- Databases
- Data Files
- Unstructured Data Sources
- Data Science
- Logical Data Models
- Physical Data Models
- XML Schema Definitions
- Master Data Management
- Applications
- Files
- Stored Procedure Definitions

Data exploration
In IBM Cloud Private for Data

Glossary and Governance

- Term (54)
- Category (4)
- Information Governance Rule (125)
- Information Governance Policy (80)
- Collection
- Label (6)

Databases

- Host (2)
- Database (2)
- Database Schema (7)
- Database Table (17)
- View (288)
- Database Column (4593)

>Analyze Data

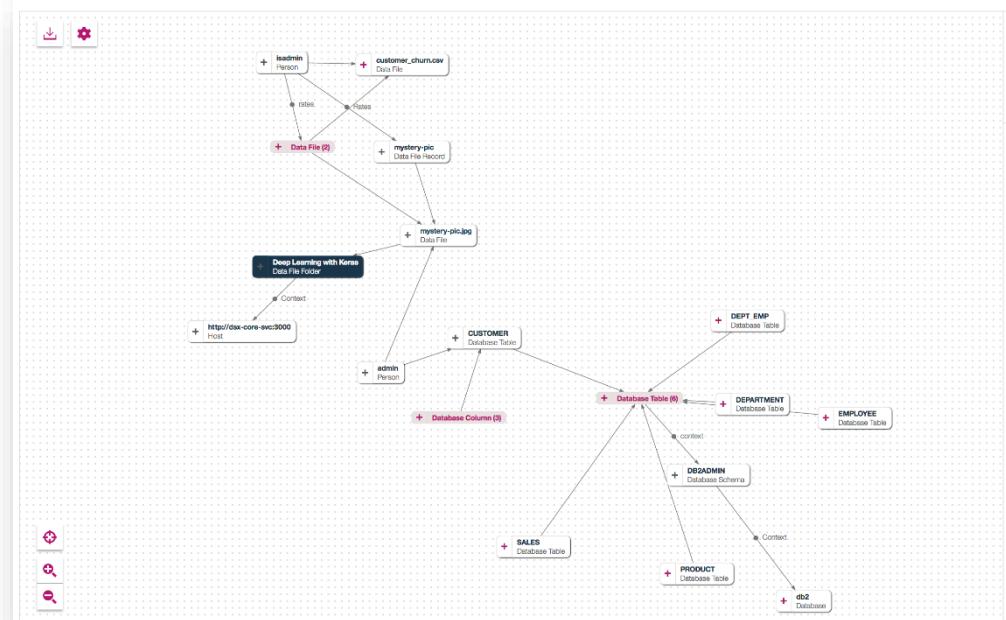
```

graph TD
    CC[Customer Churn Category] -- Context --> NGT[Net Realized Gains Term]
    NGT -- Assigned --> CACT[CUSTOMER_ACTIVITY Database Table]
    NGT -- Context --> CACT
  
```

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

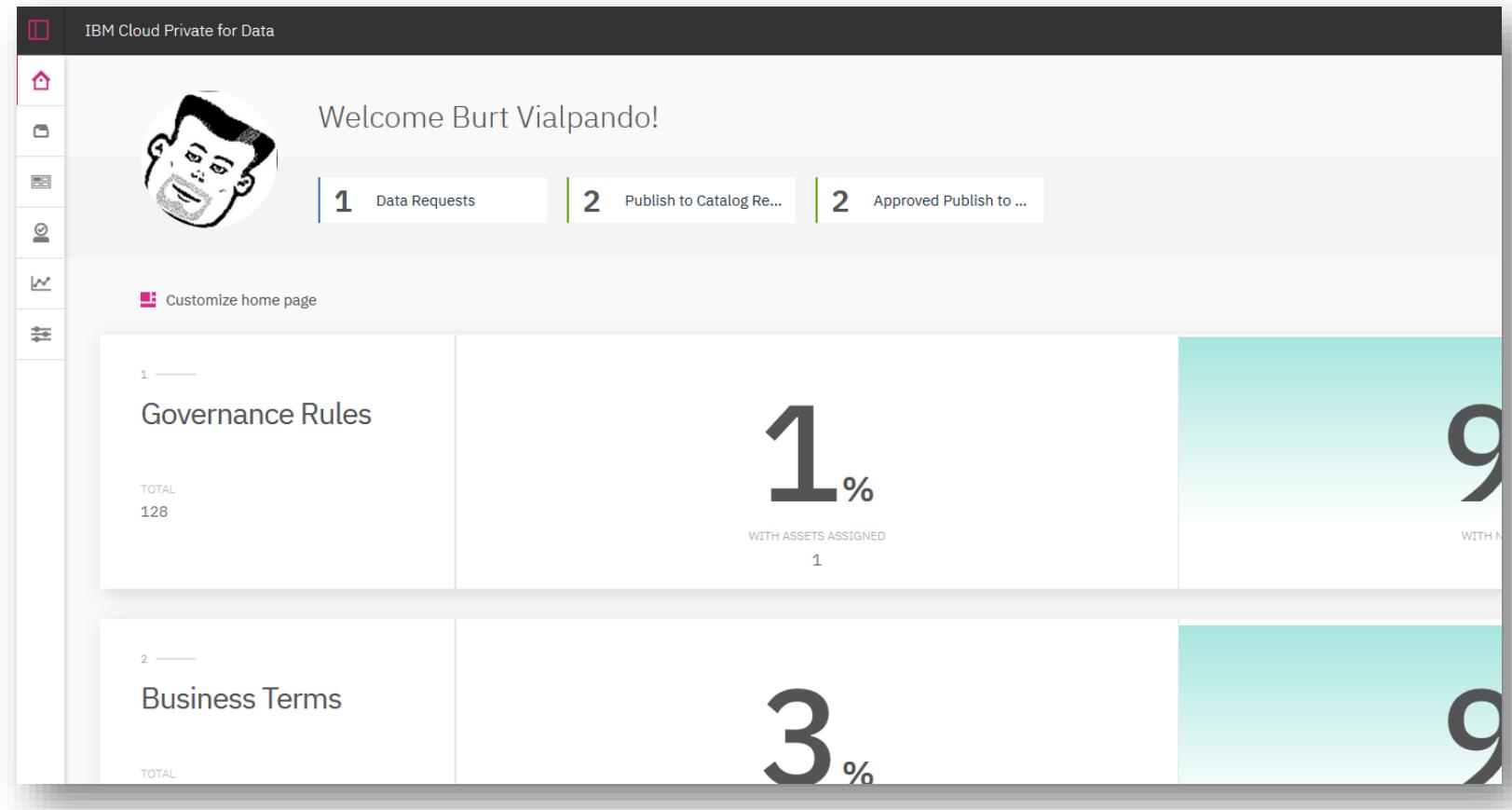


Explore deeper to understand context and usage patterns

CPD Organize

Curation dashboard – Govern and approve asset requests

- Manage asset promotion and govern quality across the organization
- Approve, reject, or ask for additional information
- Single approval process across data and analytic asset types
 - e.g. datasets, models, notebooks



CPD Organize

Approval process to make searchable

The screenshot illustrates the approval process for publishing assets to the catalog in IBM Cloud Private for Data. It shows two views of the interface: the top view displays pending publish requests, and the bottom view shows a successful publication message.

Pending Publish to Catalog Requests:

NAME	ASSET TYPE	PROJECT	OWNER	LAST UPDATED	STATUS
ICPD Modeler Demo - Vialpando Naive Bayes ML	models	ICP4DATA demo	CTP6	4 Jul 2018, 7:00 PM	Pending
GoSales Naive Bayes data.csv	datasets	ICP4DATA demo	CTP6	4 Jul 2018, 7:00 PM	

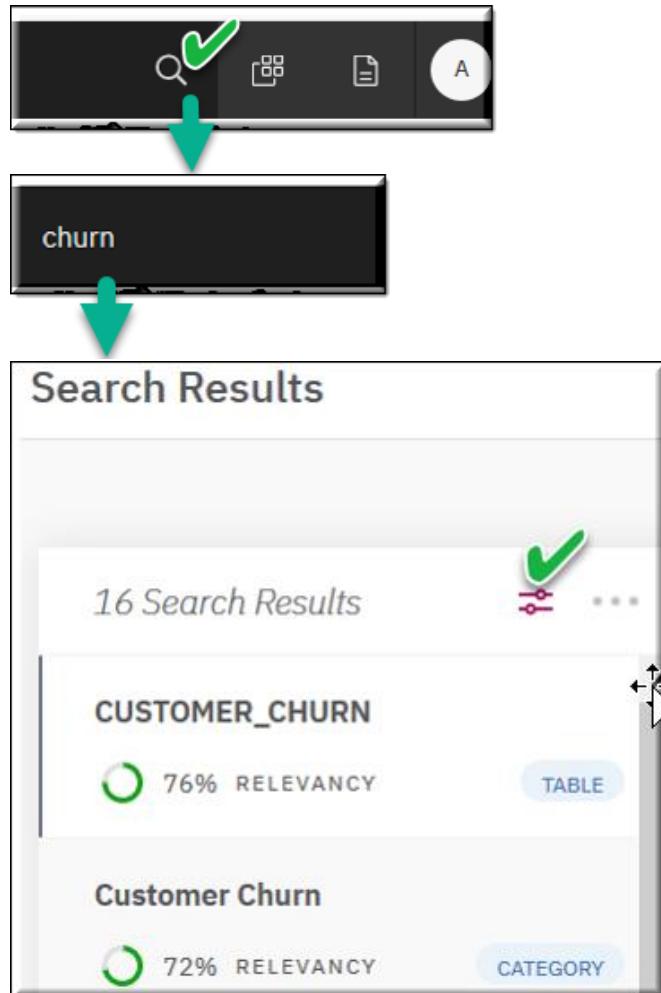
A context menu is open over the first row, listing options: Approve (with a checkmark icon), Reject, and Request information.

Successful Publication Message:

Successfully published asset ICPD Modeler Demo - Vialpando Naive Bayes ML.

CPD Organize

Shop for data – Using search



The relevancy of search results depends on the following factors:

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

CPD Organize

Transform and migrate data – Build and execute ETL jobs at scale

The screenshot shows the CPD Organize interface with the following details:

- Project:** DataClick
- JOB:** DRSToS3
- Connections:** Various data sources listed in the palette include: Connection, Amazon S3, DB2, File, Greenplum, Hive, JDBC, Kafka, Netezza, ODBC, Oracle, BDTS, DataSet, Sequential File, Teradata, and WebSphere MQ.
- Table Definitions:** Not explicitly shown in the screenshot, but implied by the connection icons.
- Jobs:** The main workspace displays a data flow graph with nodes: CUSTOMER_DEM..., CUSTOMER_ACT..., Join_15, JOBC_31, Join_20, and CUSTOMER_CHU... connected by various arrows.
- Stages:** A sidebar palette lists transformation stages: Annotation, Aggregator, Compress, Copy, Decode, Encode, Expand, Filter, Funnel, Head, Join, Lookup, Merge, Peek, Remove Duplicates, Row Generator, Sort, Tail, and Transformer.

Use powerful transformation capabilities on your data when needed.

Includes “Smart Stage Suggestion,” “Automatic metadata propagation” and other helpful autonomic tools to accelerate your tasks.

Collect and Organize

Lab 03, Lab 04 & Lab 05

<ul style="list-style-type: none">• Introduction and Setup• Executive Demo	<ul style="list-style-type: none">• Lab 01• Lab 02
<ul style="list-style-type: none">• Collect Part 1 – Connect• Organize• Collect Part 2 – Virtualize	<ul style="list-style-type: none">• Lab 03• Lab 04• Lab 05
<ul style="list-style-type: none">• Analyze Part 1 – Dashboards (optional)• Analyze Part 2 – Model Creation	<ul style="list-style-type: none">• Lab 06• Lab 07
<ul style="list-style-type: none">• Deploy and Infuse• Wrap-up	<ul style="list-style-type: none">• Lab 08• Lab 09



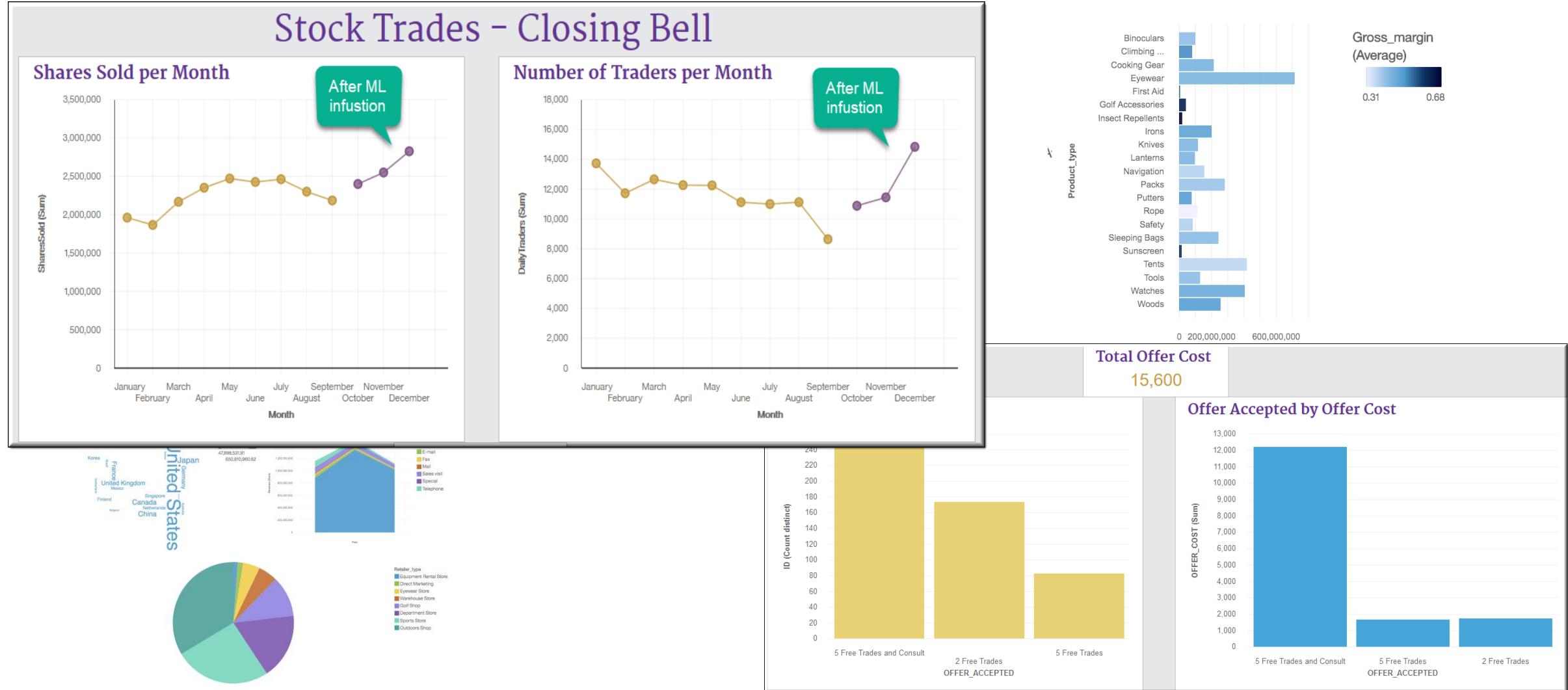
Analyze

Lab 06 – Analyze Part 1 – Dashboards (optional)

Lab 07 – Analyze Part 2 - Model Creation

Analyze Data

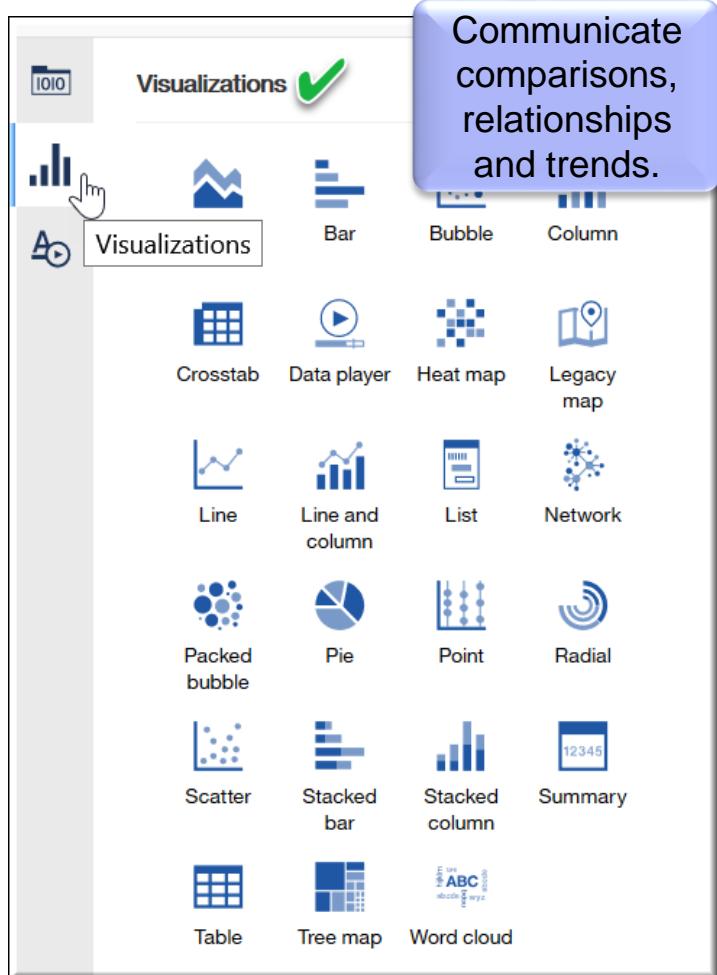
Built-in Cognos Dashboards for analytic dashboarding capabilities



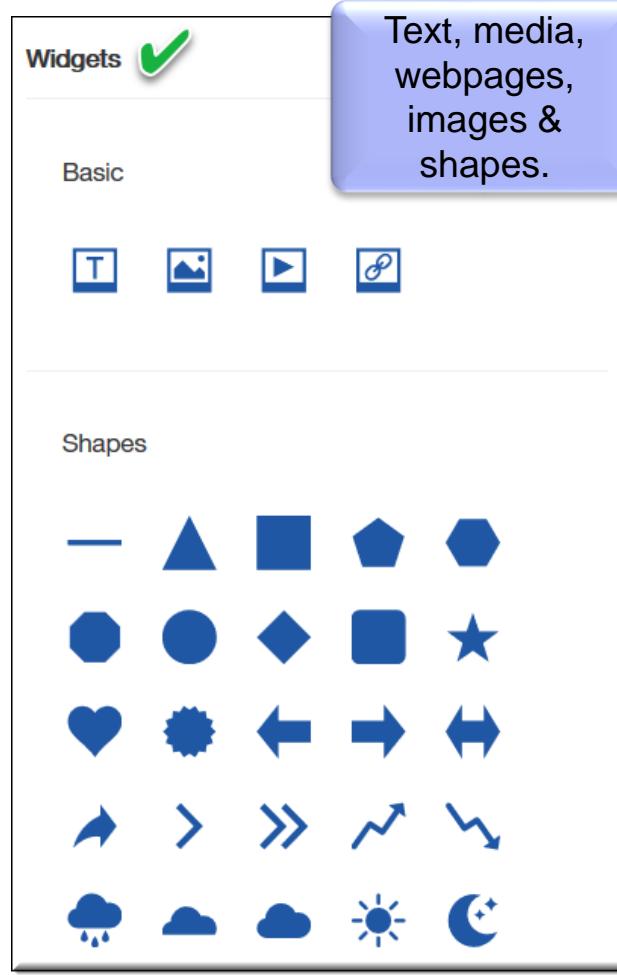
Analyze Data

Analytic dashboarding – some key capabilities

Visualizations



Widgets



Data options

The screenshot shows a list of data assets under the 'Local' tab. A callout bubble highlights the 'Local' icon with the text: 'Use data from more than one asset in your dashboard.'

Category	Icon	Name
Local	CSV icon	batchscoreresults.csv
	ML icon	TradingCustomerSparkMLE...
Remote	Cloud icon	Insert to dashboard
	Cloud icon	Insert to dashboard
Other	Cloud icon	Insert to dashboard
	Cloud icon	Insert to dashboard

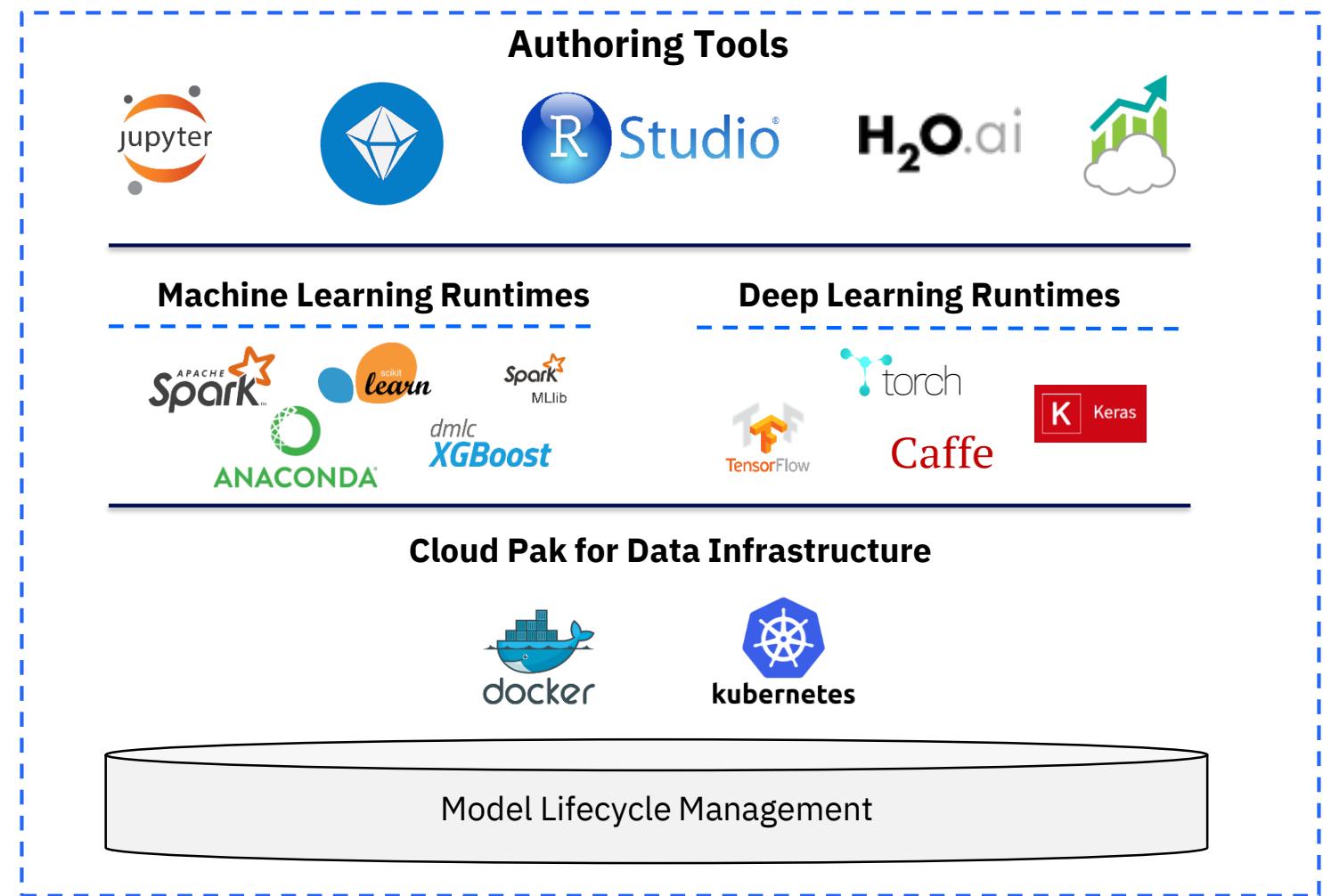
The screenshot also shows a detailed view of a data source named 'CustomerMerged'. It lists various fields and their types, along with actions like 'Filter...', 'Create calculation...', 'Create navigation path...', and 'Properties'.

Field	Type	Action
Navigation paths		+
MERGED_DEMOGRAPHIC...ADING_CUSTOMER		...
# AGE		Filter...
# CHILDREN		
DAYSSINCELASTTRADE		
ESTINCOME		
ID		
LARGESTSINGLEARTRANSACTION		

Analyze Data

Built in Watson Studio – provides open extensible data science tooling

- ✓ Best-of-breed tooling from open ecosystem
 - Authoring tools
 - Machine learning, deep learning, optimization
 - Customize environments, packages & images
- ✓ Coding and visual modeling options
- ✓ Infrastructure
 - Container-based resource management
 - Scale with distributed and GPU support
- ✓ Model Lifecycle Management
 - Dev -> Test -> Staging -> Prod
 - Versioning, release, SLAs, rolling upgrades



Analyze Data

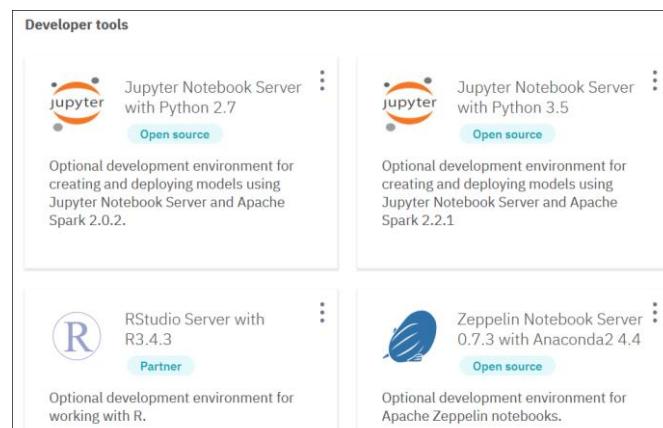
Utilize the various data science IDE's

The default environment:

- Jupyter Notebook Server 5.7.0
- Anaconda3 5.2 with the conda-forge channel
- Python 3.6
- Apache Spark 2.3.2

Environment cartridges:

- Jupyter Notebook Server with Python 2.7 or Python 3.5
- Zeppelin Notebook Server 0.7.3 with Anaconda2 4.4
- RStudio Server with R3.4.3



The screenshot shows a Jupyter Notebook interface with the title "TradingCustomerChurn > Notebooks > 01TradingCustomerChurnClassifierSparkML". The code cell contains the following R code:

```

mean(1:50)
[1] 25.5
mean(1:250)
[1] 125.5
ggplot2:::diamonds
# A tibble: 53,940 × 10
  carat cut color clarity depth table price x y z
  <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl>
1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43
2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31
3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31
4 0.29 Premium I VS2 62.4 58 334 4.28 4.23 2.63
5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75
6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48
7 0.24 Very Good I VS1 62.3 57 336 3.95 3.98 2.47
8 0.26 Very Good H SI1 61.9 55 337 4.07 4.11 2.53
9 0.22 Fair E VS2 65.1 61 337 3.87 3.78 2.49
10 0.23 Very Good H VS1 59.4 61 338 4.00 4.05 2.39
# ... with 53,930 more rows
mean(1:50)
[1] 25.5
mean(1:250)
[1] 125.5
ggplot2:::diamonds
# A tibble: 53,940 × 10
  carat cut color clarity depth table price x y z
  <dbl> <ord> <ord> <ord> <dbl> <dbl> <dbl> <dbl>
1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43
2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31
3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31
4 0.29 Premium I VS2 62.4 58 334 4.28 4.23 2.63
5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75
6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48
7 0.24 Very Good I VS1 62.3 57 336 3.95 3.98 2.47
8 0.26 Very Good H SI1 61.9 55 337 4.07 4.11 2.53
9 0.22 Fair E VS2 65.1 61 337 3.87 3.78 2.49
10 0.23 Very Good H VS1 59.4 61 338 4.00 4.05 2.39
# ... with 53,930 more rows

```

The output cell shows the results of the mean() function applied to a dataset, resulting in values like [1] 25.5 and [1] 125.5.

Analyze Data

Utilize the various algorithms

Random Forest

4. Build SparkML Random Forest classification model

[Top](#)

We instantiate a decision-tree based classification algorithm, namely, RandomForestClassifier. Next we define a pipeline. MLlib standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline.

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

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

# Convert indexed labels back to original labels.
labelConverter = IndexToString(inputCol="prediction", outputCol="predictedLabel", labels=[0, 1])
```

ID	CHURNRISK	label	predictedLabel	prediction	probability
0	4	High	0.0	High	[0.8735308585578527, 0.0007194244604316547, 0.0...
1	7	High	0.0	High	[0.7860462153640373, 0.1325810995942404, 0.081...
2	8	Medium	2.0	High	[0.5341352413075571, 0.00030864197530864197, 0...
3	9	Medium	2.0	Medium	[0.2145249547966675, 0.007092436847936796, 0.7...
4	15	Low	1.0	Low	[0.17070740206723437, 0.7844143642467721, 0.04...
5	18				[0.145993580858905, 0.7876603588435366, 0.06...

```
print('Model Precision = {:.2f}.'.
```

Model Precision = 0.92.

Overall Statistics
Model F-measure = 0.917741935483871

Statistics by Class
Class 0.0 F-Measure = 0.9265905383360522
Class 1.0 F-Measure = 0.9864253393665159
Class 2.0 F-Measure = 0.7243243243243243

Evaluation

Naïve Bayes

```
nb=NaiveBayes(labelCol="label", featuresCol="features")

stages_nb = stages
stages_nb[-2] = nb

pipeline_nb = Pipeline(stages = stages_nb)

# Build models
model_nb = pipeline_nb.fit(train)
results_nb = model_nb.transform(test)

print('Naive Bayes Model Precision = {:.2f}'.format(results_nb.filter(
```

Naive Bayes Model Precision = 0.51.

Analyze Data Build Data Science & Machine Learning models

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

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

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

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

pipeline = Pipeline(stages = stages)

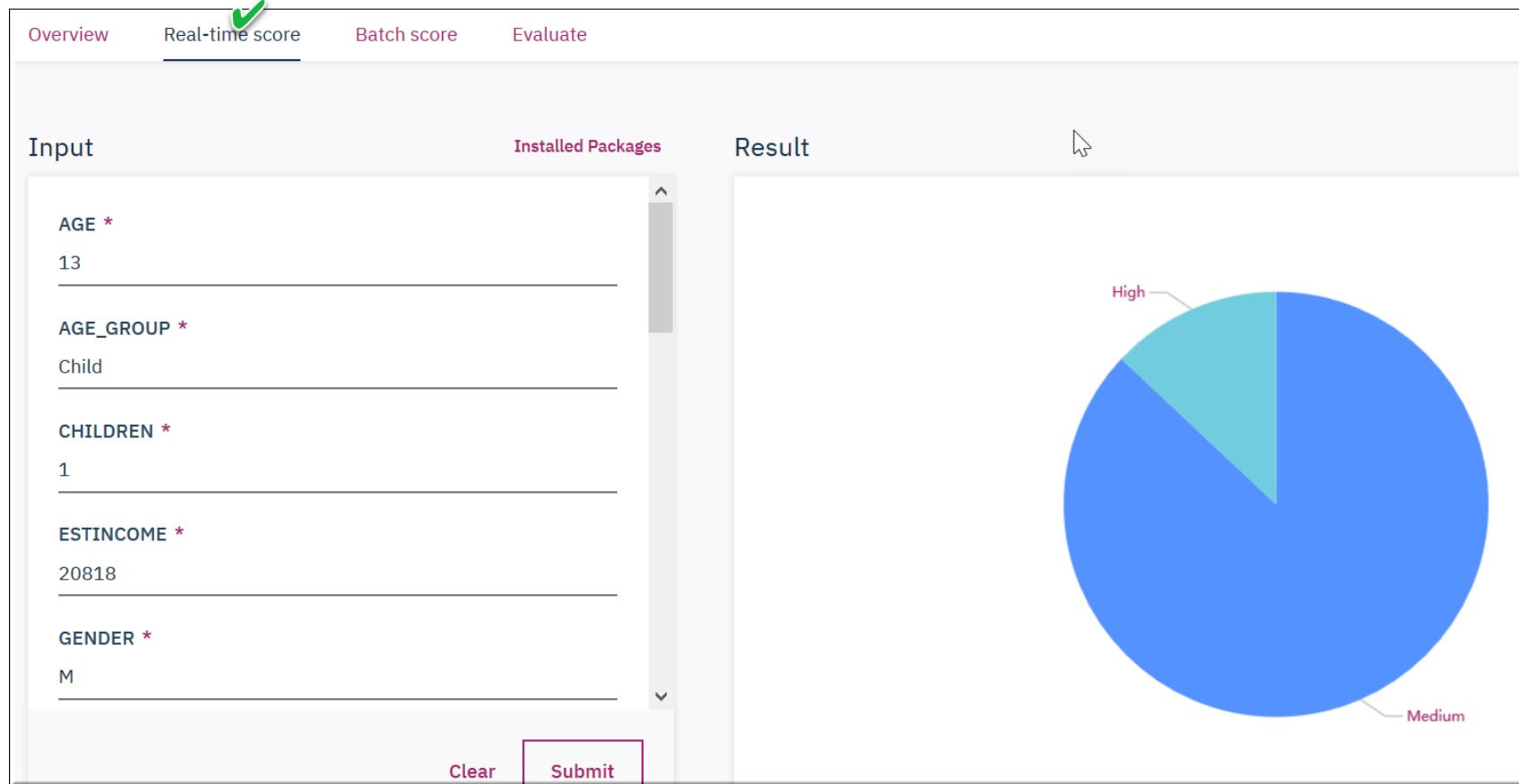
In [68]: # Split data into train and test datasets
train, test = df_churn.randomSplit([0.7,0.3], seed=100)
train.cache()
test.cache()
```



- ✓ 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 CICD 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 Data

Model scoring in real-time – one set of values at a time



The following machine learning model types are supported for real-time, batch scoring and evaluation:

- Spark ML
- PMML with online scoring
- Custom models with batch scoring
- scikit-learn 0.19.1 (Python 2.7 and Python 3.5) - 0.19.1 (GPU-Python 3.5) with pickle or joblib format
- XGBoost 0.7.post3 (Python 2.7 and 3.5) - 0.71 (GPU-Python 3.5)
- Keras 2.1.3 (Python 2.7 and Python 3.5) - 2.1.5 (GPU-Python 3.5)
- TensorFlow 1.5.0 (Python 2.7 and Python 3.5) - 1.4.1 (GPU-Python 3.5)
- WML

Analyze Data

Model scoring in batch – many sets of values at a time

Overview Real-time score **Batch score** Evaluate

Batch scoring script inputs

Spark cluster *

Local Spark

Input data set *

batchscoreresults.csv

Output data set *

Local file Remote data set

batchscoreresultsout.csv

26

Result

```

1 #!/usr/bin/python
2
3 import sys, os
4 from pyspark.sql import SparkSession
5 from pyspark.ml import Pipeline, Model, PipelineModel
6 from pyspark.sql import SQLContext
7 import pandas
8 import dsx_core_utils, re, jaydebeapi
9 from sqlalchemy import *
10 from sqlalchemy.types import String, Boolean
11
12
13 # setup dsxr environmental vars from command line input
14 from dsx_ml.ml import dsxr_setup_environment
15 dsxr_setup_environment()
16
17 # define variables
18 args = {'execution_type': 'DSX', 'target': '/datasets/batchscoreresultsout.csv', 'source': '/datasets/batchscoreresults.csv',
19 input_data = os.getenv("DEF_DSX_DATASOURCE_INPUT_FILE", (os.getenv("DSX_PROJECT_DIR") + args.get("source")))
20 output_data = os.getenv("DEF_DSX_DATASOURCE_OUTPUT_FILE", (os.getenv("DSX_PROJECT_DIR") + args.get("target")))
21 model_path = os.getenv("DSX_PROJECT_DIR") + os.path.join("/models", os.getenv("DSX_MODEL_NAME", "TradingChurnRiskClassification"))
22
23 # create spark context
24 spark = SparkSession.builder.getOrCreate()
25 sc = spark.sparkContext
26

```

Generate batch script Run now

“Scoring” (or “prediction”) produces an outcome given a set of input values against a trained model.

Analyze Data

Model evaluation

The screenshot shows the 'Evaluate' tab selected in the top navigation bar. The left panel, titled 'Model evaluation script inputs', contains the following configuration:

- Spark cluster ***: Local Spark
- Input data set ***: TradingCustomerSparkMLEval.csv
- Evaluator ***: Multiclass
- Threshold metric ***: F1 Score
- Threshold ***: A slider set between 0 and 1, with markers at Min: 0.30 and Mid: 0.70.

At the bottom of this panel is a red-bordered button labeled 'Generate evaluation script'.

The right panel, titled 'Result', displays a scrollable code editor containing a Python script. The script imports various libraries like pandas, json, uuid, requests, and PySpark, and defines variables and logic for model evaluation. It includes imports from the dsx_ml.ml module, such as save_evaluation_metrics and dsxr_setup_environment.

At the bottom right of the result panel is a red-bordered button labeled 'Run now'.

A callout bubble on the right side of the interface states: "Evaluation quantifies the quality of a model's overall scoring output."

Analyze Data

Model overview and tracking with accuracy, features, label columns, scripts, etc.

TradingChurnRiskClassificationSparkML v1

This is a SparkML Model to Classify Trading Customer Churn Risk

LAST MODIFIED
28 May 2019, 9:33 AM

TYPE
Spark

ALGORITHM
PipelineModel (Classification)

Overview Real-time score Batch score Evaluate

Accuracy

94%

Accuracy history

Features

Name	Type
AGE	int
AGE_GROUP	str

Label columns

Name
CHURNRISK

Analyze Data

Premium cartridge: Watson Studio SPSS Modeler

The screenshot shows the IBM Watson Studio interface with the SPSS Modeler cartridge selected. The main workspace displays a data flow diagram for a project titled "Chronic Kidney Disease - SPSS Modeler". The flow starts with a "UCI ML Repository" node, followed by a "Filter" node, a "Select" node, and a "Type" node. These are connected to a "K-Means" node. From the "K-Means" node, the flow splits into two parallel paths. The left path leads to a "Partition" node, which then connects to a "Target Di..." node. The right path leads to a "Decision ..." node, which then connects to a "Cluster D..." node. Both the "Partition" and "Decision ..." nodes lead to a "Table" node. Finally, the "Table" node connects to an "Analysis" node. A "Data Audit" node is also present in the diagram. A context menu is open over the "K-Means" node, listing options such as Open, Disconnect, Preview, Edit, Delete, Create supernode, Save branch as a model, and Run.

Watson Studio 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.

MLP Neural Network Model

Network Diagram

Network Diagram

Network Diagram

Analyze

Lab 06 & Lab 07

<ul style="list-style-type: none">• Introduction and Setup• Executive Demo• Collect Part 1 – Connect• Organize• Collect Part 2 – Virtualize	<ul style="list-style-type: none">• Lab 01• Lab 02• Lab 03• Lab 04• Lab 05
<ul style="list-style-type: none">• Analyze Part 1 – Dashboards (optional)• Analyze Part 2 – Model Creation	<ul style="list-style-type: none">• Lab 06• Lab 07
<ul style="list-style-type: none">• Deploy and Infuse• Wrap-up	<ul style="list-style-type: none">• Lab 08• Lab 09



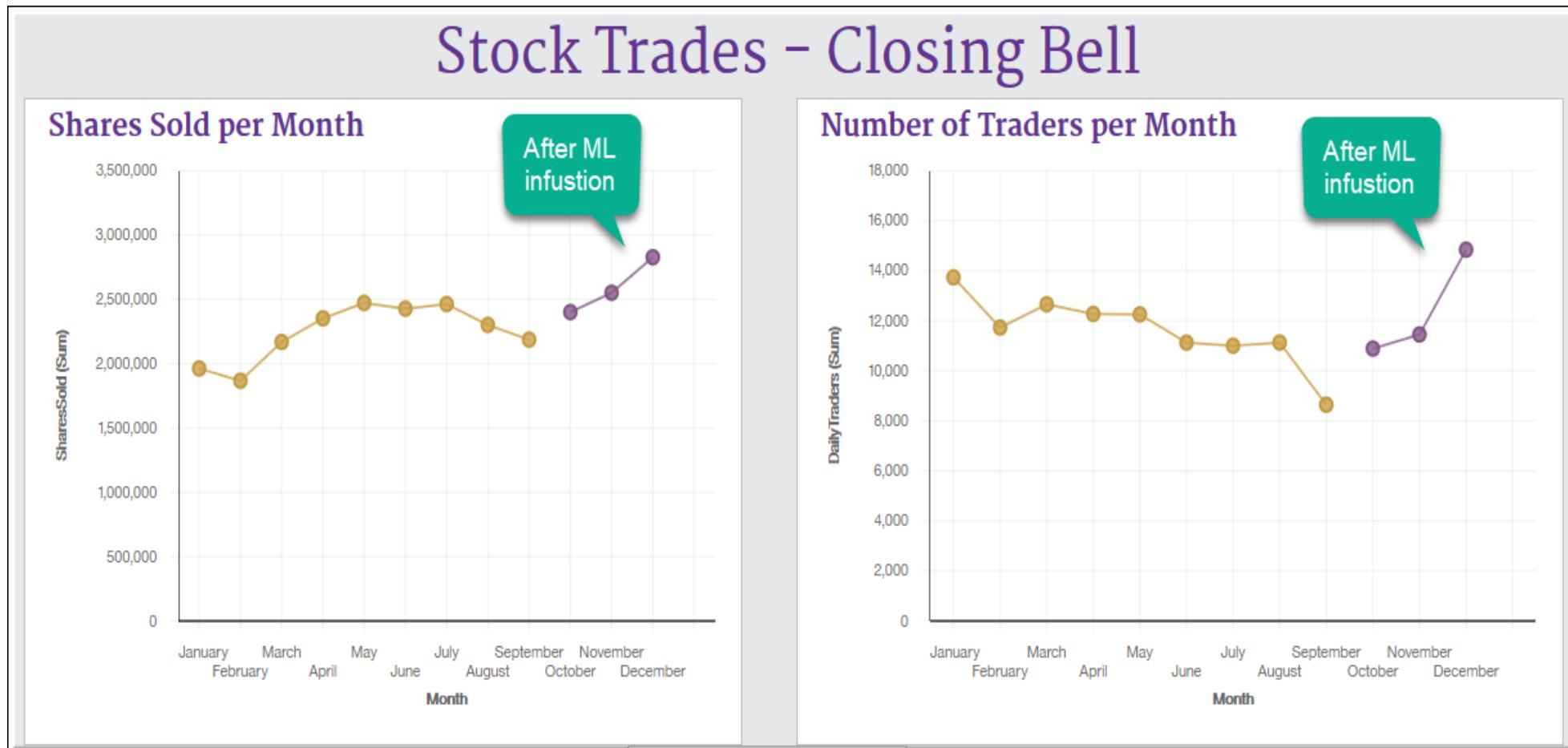
Deploy and Infuse

Lab 08 Deploy and Infuse

Cloud Pak for Data

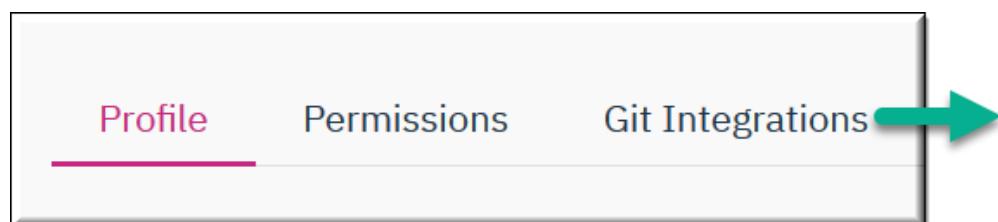
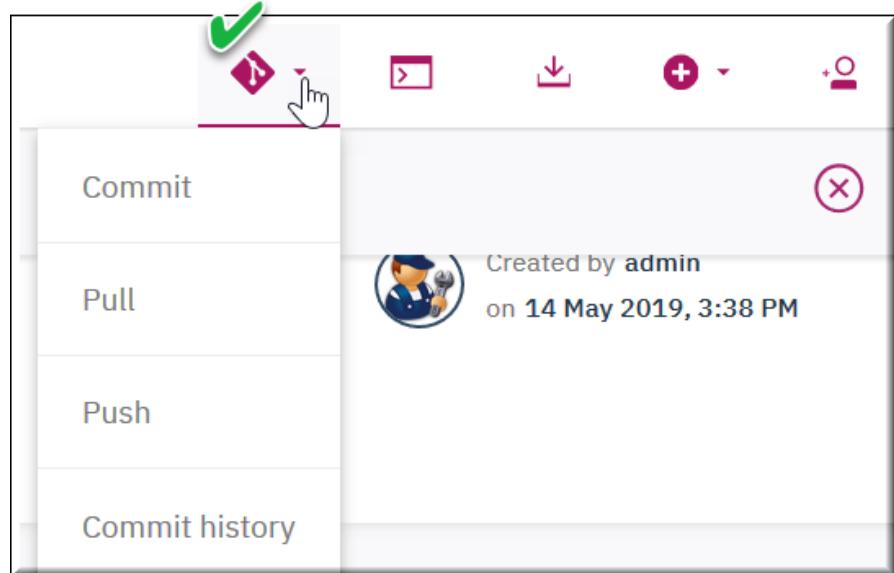
Impact to an application with an infused ML model

Assess the Impact – Continuous Improvement helps the bottom line



CPD Deploy - Model Management and Deployment

Integration with source control



- ✓ Integration with GitHub, GitHub Enterprise, BitBucket, and BitBucket Server
- ✓ Model Deployment – a self-service application that takes the model and provides REST API endpoints
- ✓ Harvest insight from data in a repetitive fashion by integration to DevOps
- ✓ CPD platform automates these processes to cut short significantly on time

A screenshot of the CPD Deploy interface showing the "Git Integrations" tab selected. The tabs at the top are Profile, Permissions, and Git Integrations (selected). Below the tabs, there's a section titled "Add token" with instructions: "Visit [GitHub personal access tokens](#), select repository scope and generate a token." There's also a "Platform*" label with radio buttons for GitHub (selected), GitHub Enterprise, BitBucket, and BitBucket Server.

CPD Deploy - Model Management and Deployment Project Releases

The diagram illustrates the deployment process from a project release to the assets dashboard.

Project releases (Left Panel):

- stocktrader** (Project Name)
- MEMBERS**: Placeholder icon
- SOURCE**: TradingCustomerChurn
- DEPLOYMENTS**: 0
- LAST MODIFIED**: 30 May 2019, 12:56 PM

Assets Dashboard (Right Panel):

- Dashboard**, **Deployments**, **Assets** (Selected), **Data Sources**
- Search by asset name** input field
- All** (selected)
- Flows**
- Models** (selected)

A green arrow points from the Project releases panel to the Assets tab in the dashboard. A green checkmark is placed on the **Assets** tab. A green arrow points from the Models section in the dashboard to the deployment details below.

NAME	ASSET	TYPE	VISIBILITY	DATE STARTED	AVAILABILITY
stocktrader	TradingChurnRiskClassificationSparkML v2	Web service	-	30 May 2019, 1:07 PM	Enabled

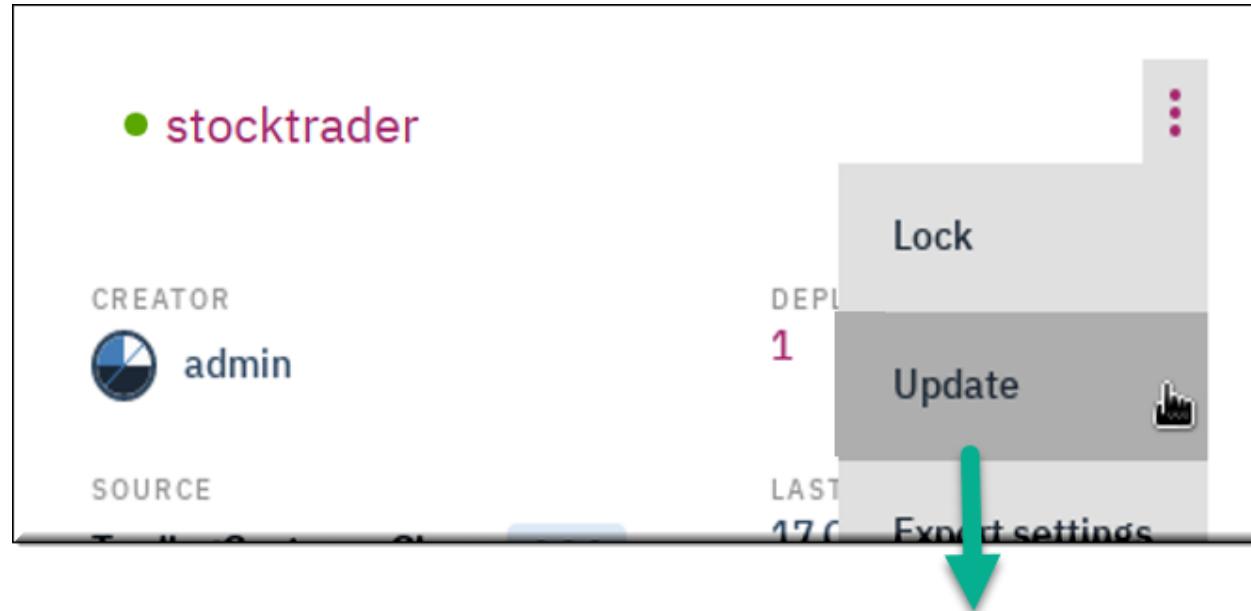
CPD Deploy - Model Management and Deployment

Project Release deployable assets

Asset	Job	Web service	App	Comments
Notebooks	✓		✓	Only Jupyter notebooks can be deployed. Zeppelin is not supported.
Models		✓		Requires that you associate a batch scoring script with it.
R Shiny			✓	Deploys the R code inside of an R pod.
Flows	✓			Flows from SPSS Modeler.
Decision Optimization models		✓		Provides a REST API to submit and execute optimization jobs.
Scripts	✓	✓		

CPD Deploy - Model Management and Deployment

Project Releases updates – deploy model v2, v3, etc.



Deploy Model

Tooling to decrease operational deployment times

- ✓ Update existing releases to add new versions of the models
- ✓ Run multiple versions of the same model and easily scale them up and down
- ✓ Scheduling of pods is fully automatic – self-service without a need of IT staff

NAME	ASSET	TYPE	VISIBILITY	DATE STARTED	AVAILABILITY
stocktrader	TradingChurnRiskClassificationSparkML v2	Web service	—	30 May 2019, 1:07 PM	✓ Enabled

CPD Deploy - Model Management and Deployment

Easily score models with an API or a generated curl command

The screenshot illustrates the CPD Deploy interface for Model Management and Deployment. It shows the following components:

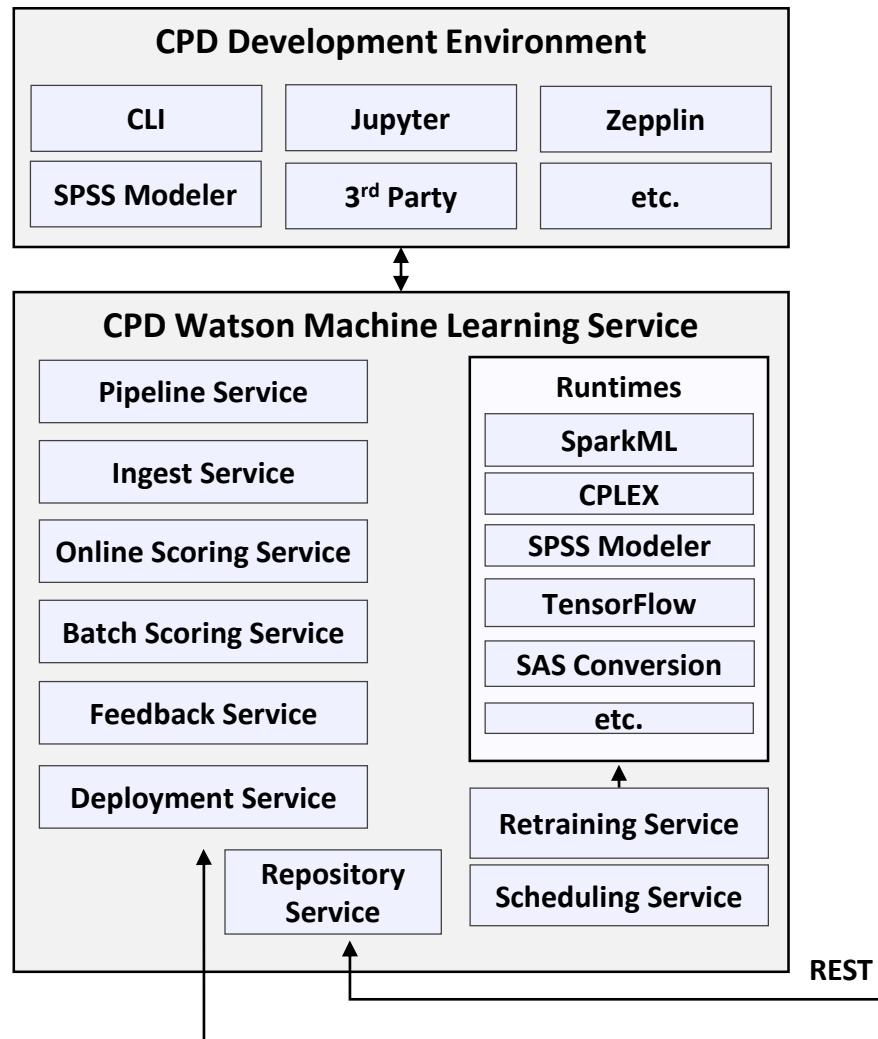
- API Tab:** Selected tab, indicated by a green checkmark.
- Request Section:** Contains fields for "Function name *" (set to "score") and "Body *". The "Body" field contains a JSON input string:

```
{"input_json": [{"ID": 718, "NetRealizedLosses_YTD": -0.795934533, "SmallestSingleTransaction": -0.1127903522, "HomeOwner_Y": 1, "Status_S": 1, "Gender_F": 0, "DaysSinceLastTrade": -1.0249341976, "HomeOwner_N": 0, "LargestSingleTransaction": -0.1127903522, "Status_D": 0, "TotalDollarValueTraded": -0.1885126001, "Children": 0.1734204213, "NetRealizedGains_YTD": 0.6503041139, "Gender_M": 1, "TotalUnitsTraded": -0.2358396702, "EstIncome": 0.6434087651, "PercentCha
```
- Response Section:** Displays the API response in JSON format:

```
1 {
2   "result": [
3     "classes": [
4       "High",
5       "Low",
6       "Medium"
7     ],
8     "probabilities": [
9       [
10        0.049672079124133905,
11        0.9037124663576185,
12        0.04661545451824756
13      ]
14    ],
15    "predictions": [
16      "Low"
17    ],
18  },
19  "stdout": [
20    "+---+-----+",
21    "| AGE|AGE_GROUP|CHILDREN|",
22    "+---+-----+",
23    "| 12| Child|
```
- Generate code:** A button with a green checkmark that generates a curl command for the API call. The generated curl command is:

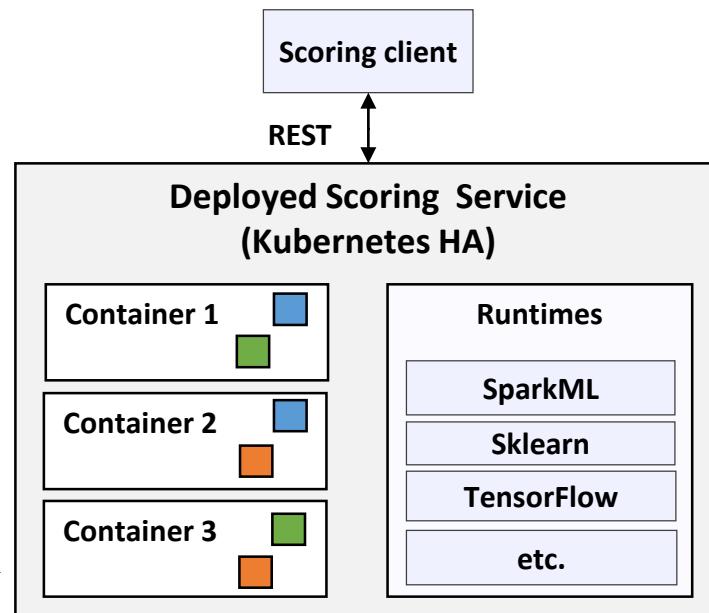
```
1 curl -k -X POST \
2   https://10.77.200.160:31843/dmodel/v1/award/pyscript/stocktrader/score \
3   -H 'Authorization: Bearer eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCJ9.eyJcI2VybmFtZSI6ImFkbWluIiwicGFja2FnZU5hbWUiOiJzdG9ja3RyYi
4   -H 'Cache-Control: no-cache' \
5   -H 'Content-Type: application/json' \
6   -d '{"args":{"input_json": [{"AGE":12,"AGE_GROUP":"Child","CHILDREN":0,"ESTINCOME":28770,"GENDER":"F","HOMEOWNER":"N","ID":718}]}'
```

CPD Deploy - Model Management and Deployment Architecture



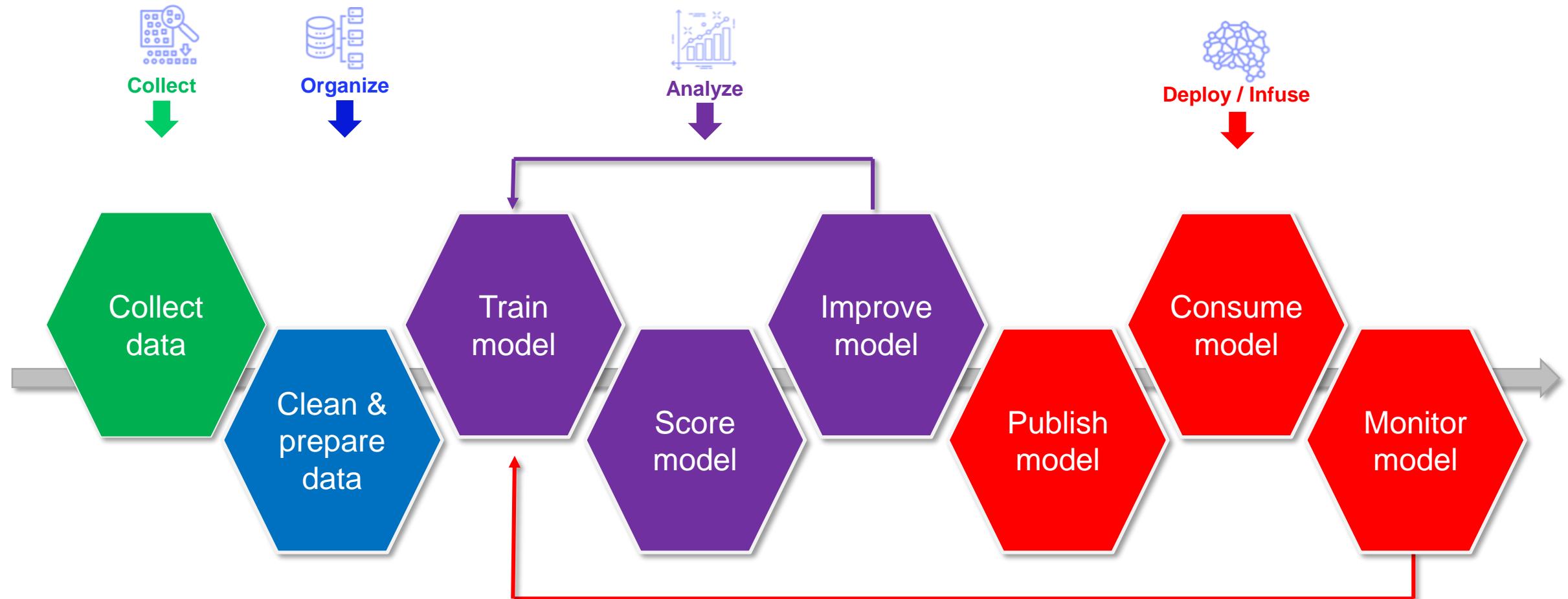
Model Management and Deployment

- CI/CD style pipelines for *AI DevOps* across non-production and production environments
- Model monitoring, with outage free rolling upgrades and rollbacks
- Set *Quality-of-Service* per model through scale-up and scale-out
- Autogenerated code for triggering on-demand or scheduled jobs



Machine Learning Model Lifecycle

CPD simplifies the entire process





Wrap-up

Lab 09 Wrap-up

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:

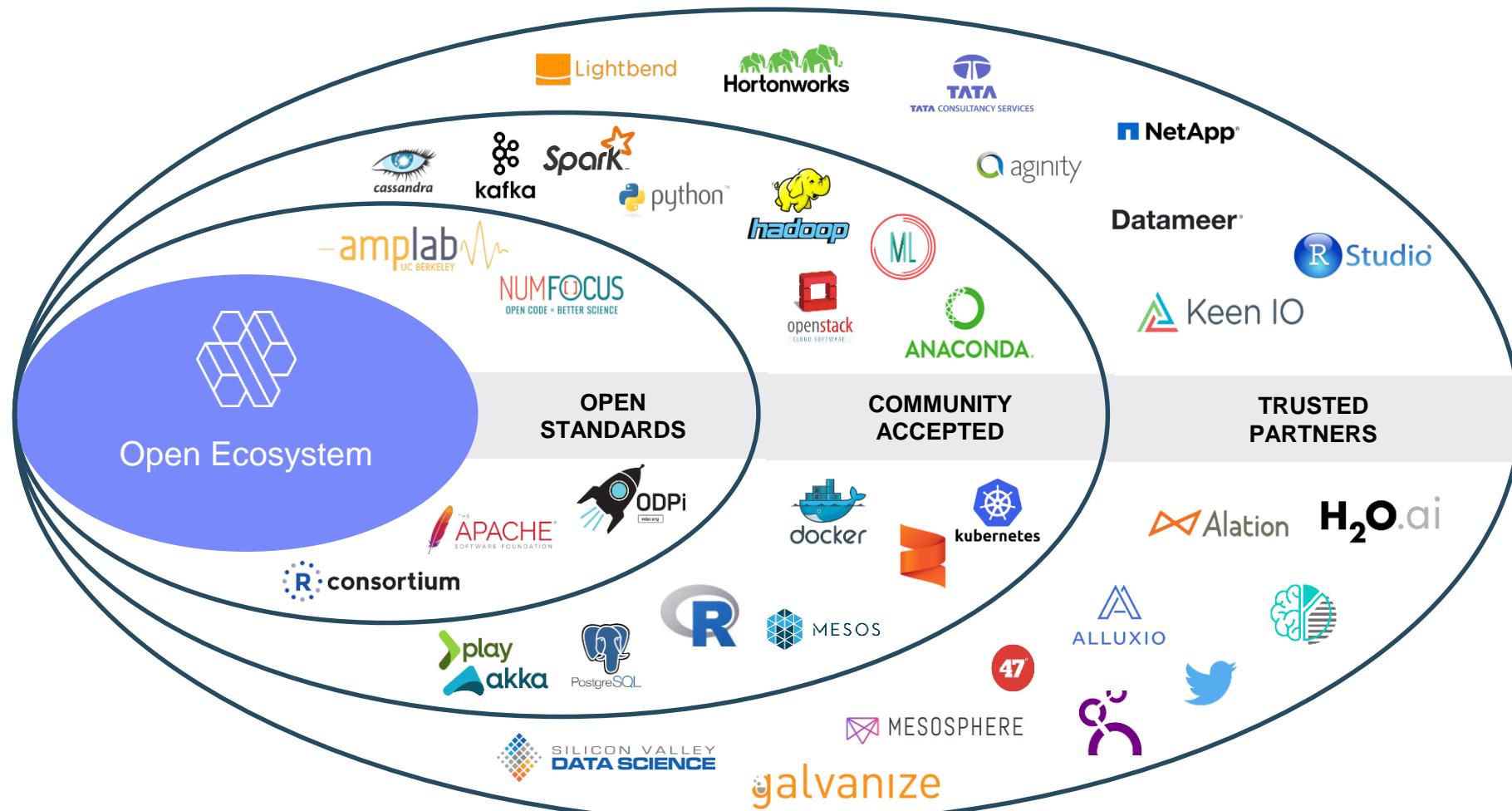
FISMA High “ready” with System Security Plans, spanning 350 controls:

- 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



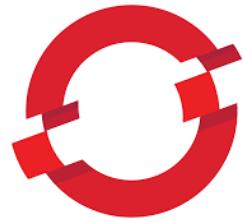
IBM's approach to Open technology: <https://developer.ibm.com/articles/cl-open-architecture-update/>

CPD Multi-cloud support

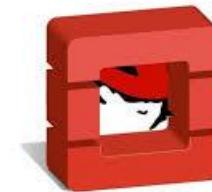
Build Once - Run anywhere
In your own data center
Or the cloud infrastructure of your choice



IBM Cloud



OPENSHIFT



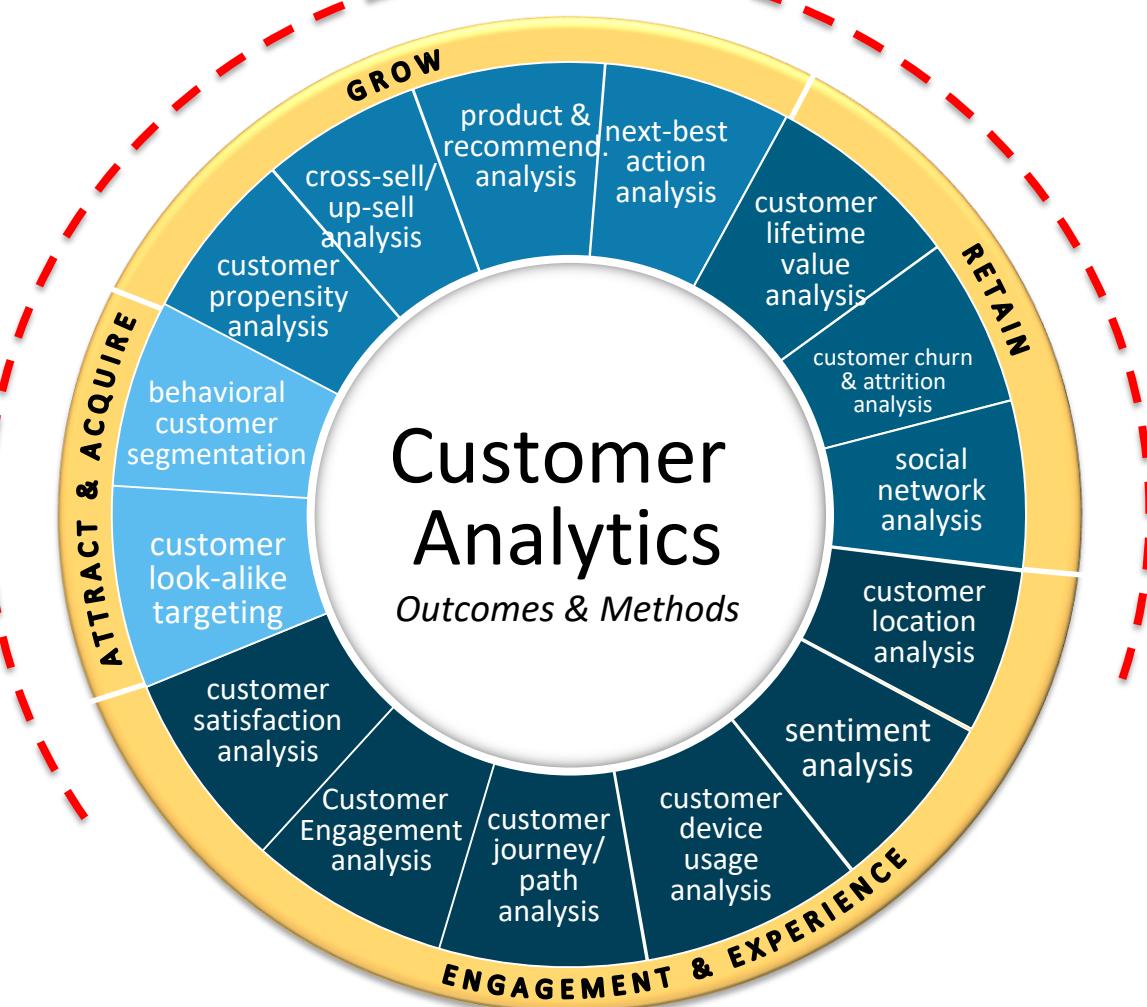
openstack™



Helps avoid vendor lock in

CPD Use Cases

Industry agnostic



Create a Customer Focused Enterprise

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

These use cases apply to most industry verticals

CPD Use Cases

Industry specific

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

Cloud Pak for Data Differentiation :

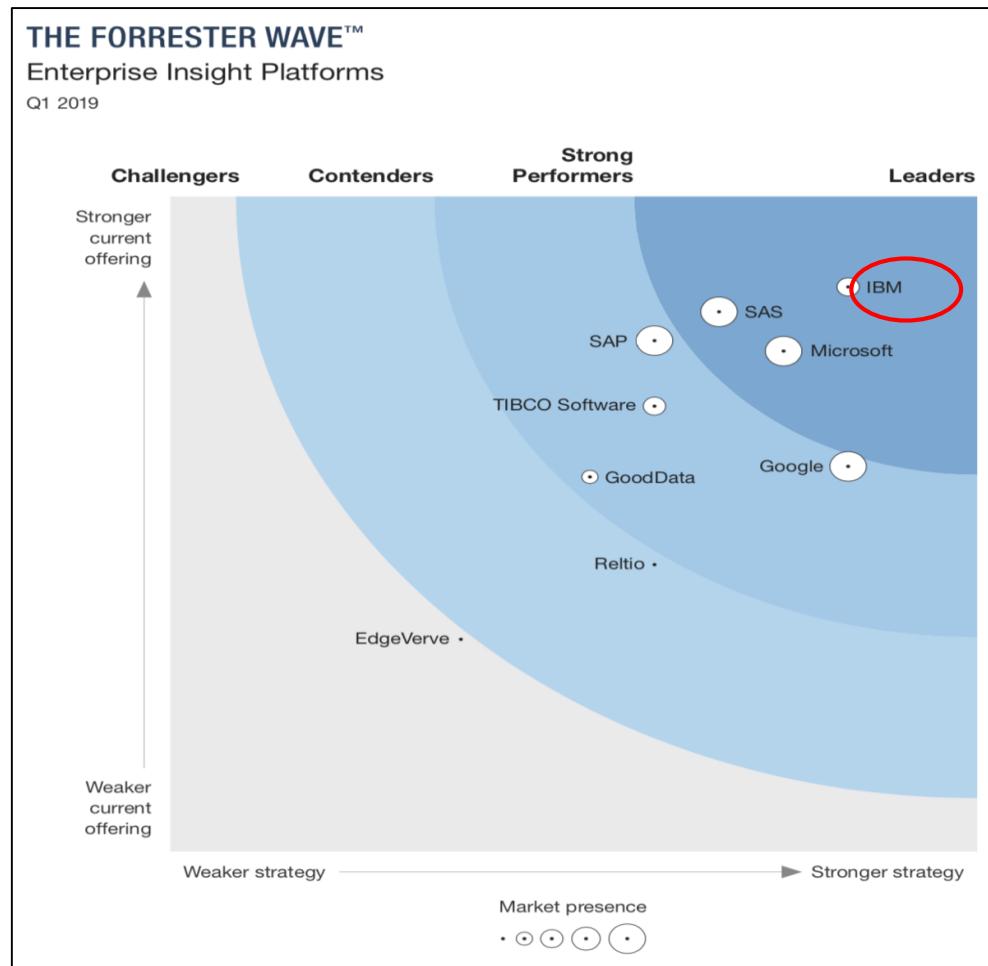
- ✓ Operationalize models in a matter of minutes (Deploy, scale & manage models with minimum effort)
- ✓ Model governance (Lineage and provenance – Who created, when, what data was used, comments, ratings etc.)



Use Case(s)	Aerospace & Defense	Automotive	Banking	Chemicals & Petroleum	Consumer	Education	Electronics	Energy, Environmental & Utilities	Financial Markets	Government	Healthcare & Life Sciences	Insurance	Industrial Products	Telco, Media & Entertainment	Travel & Transportation
Predictive Maintenance	X							X							X
Real time analytics (IOT)	X	X		X											X
Customer Churn / Retention		X	X												X
Anomaly Detection	X	X					X							X	
Regulatory Compliance			X						X				X		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”



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.

CPD Packaging Editions

Cloud Pak for Data

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

Cloud Native Edition

- Beginning with a minimum of 24 VPCs
- Expands in increments of 1 VPC
- Up to a maximum of 64 VPCs

Basic Support

Enterprise Edition

- Beginning with a recommended minimum of 48 VPCs
- Select 24 VPC configurations supported
- Expands in increments of 1 VPC
- No maximum

Basic & Premium Support

CPD Packaging & Cartridges (v2.5)

1. Cloud Pak for Data Services

Collect

Powered by: new Db2 Technology & Db2 Warehouse

Data virtualization	
PostgreSQL	Db2 Warehouse
IBM Streams	Db2 Event Store

Organize

Powered by: Information Analyzer, IGC & Data Stage

Watson Knowledge Catalog
Governance
Data Integration & Discovery

Analyze

Powered by: Cognos CDE

Data Visualization & Dashboards
Industry Accelerators: Offer Affinity, Life Event Prediction, Intelligent Maint. for Telco.

Deploy

Powered by: Watson Studio

Data Science: Model Design & Deployment
Watson OpenScale

OpenShift

2. Premium Cartridge Services (Purchase license or BYOL)

Collect

Db2 AESE
Db2 BigSQL

Organize

Infosphere DataStage for CPD
Infosphere Regulatory Accelerator
Infosphere multi-cloud Data Mgmt
Infosphere Entity Resolution

Analyze

Cognos Analytics
Watson Studio Premium
(SPSS Modeler, Model builder,
Decision Optimization, Watson Explorer
Data Refinery, Streams Designer)

Infuse

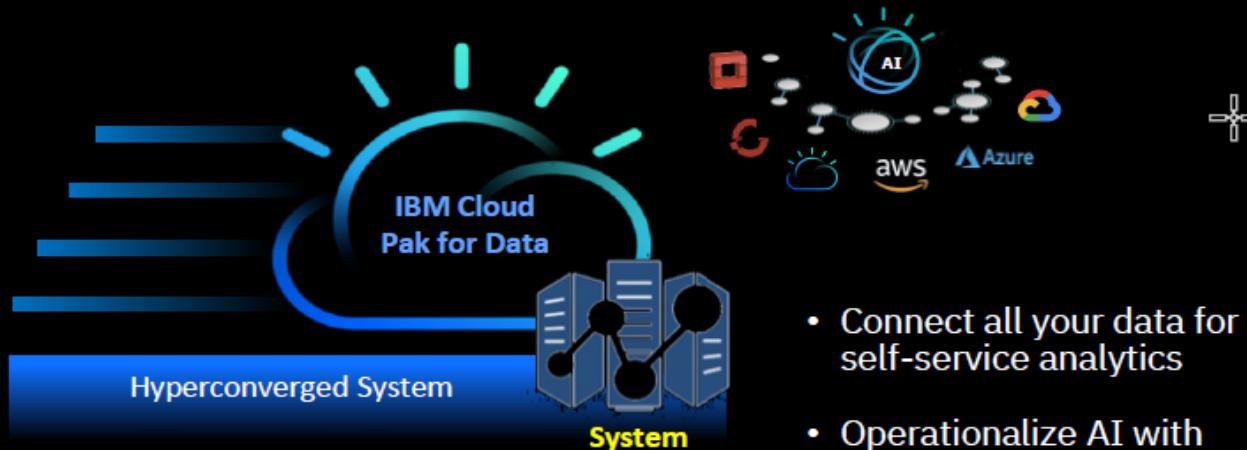
Auto AI
Watson Bundles
(Discovery, Assistant, Speech to Text,
Natural Language Understanding, API Kit,
Watson Knowledge Studio)

3. Third Party Add-Ons



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

Key differentiators (vs. Microsoft, AWS and Google)

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



Turbocharged digital transformation

Read the story in any of these magazines:

Business Chief US

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

Business Chief Canada

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

Gigabit Magazine

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



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

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

Cloud Pak for Data

Industry: Telecommunications
Geography: North America

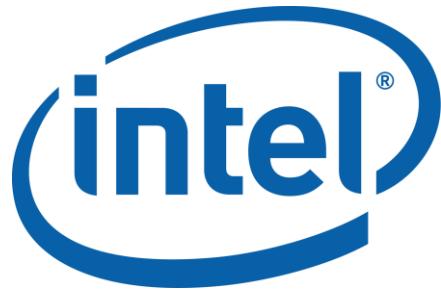
Simplify and automate how your organization turns data into insights within a unified, all-in-one design.

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

Michelle Gehl

VP Networks OSS Applications and Operations, Sprint





Is AI your priority? Start with a data strategy

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

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

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

Cloud Pak for Data

Industry: Technology

Geography: North America

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

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

Melvin Greer

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

Deploy and Infuse, Wrap-up

Lab 08 & Lab 09

<ul style="list-style-type: none">• Introduction and Setup• Executive Demo• Collect Part 1 – Connect• Organize• Collect Part 2 – Virtualize	<ul style="list-style-type: none">• Lab 01• Lab 02• Lab 03• Lab 04• Lab 05
<ul style="list-style-type: none">• Analyze Part 1 – Dashboards (optional)• Analyze Part 2 – Model Creation	<ul style="list-style-type: none">• Lab 06• Lab 07
<ul style="list-style-type: none">• Deploy and Infuse• Wrap-up	<ul style="list-style-type: none">• Lab 08• Lab 09

THANK YOU

We appreciate your feedback.
Please fill out the survey in order to improve this event.

Copyright and trademarks

© Copyright IBM Corporation 2019

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
July 2019

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

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

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

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

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