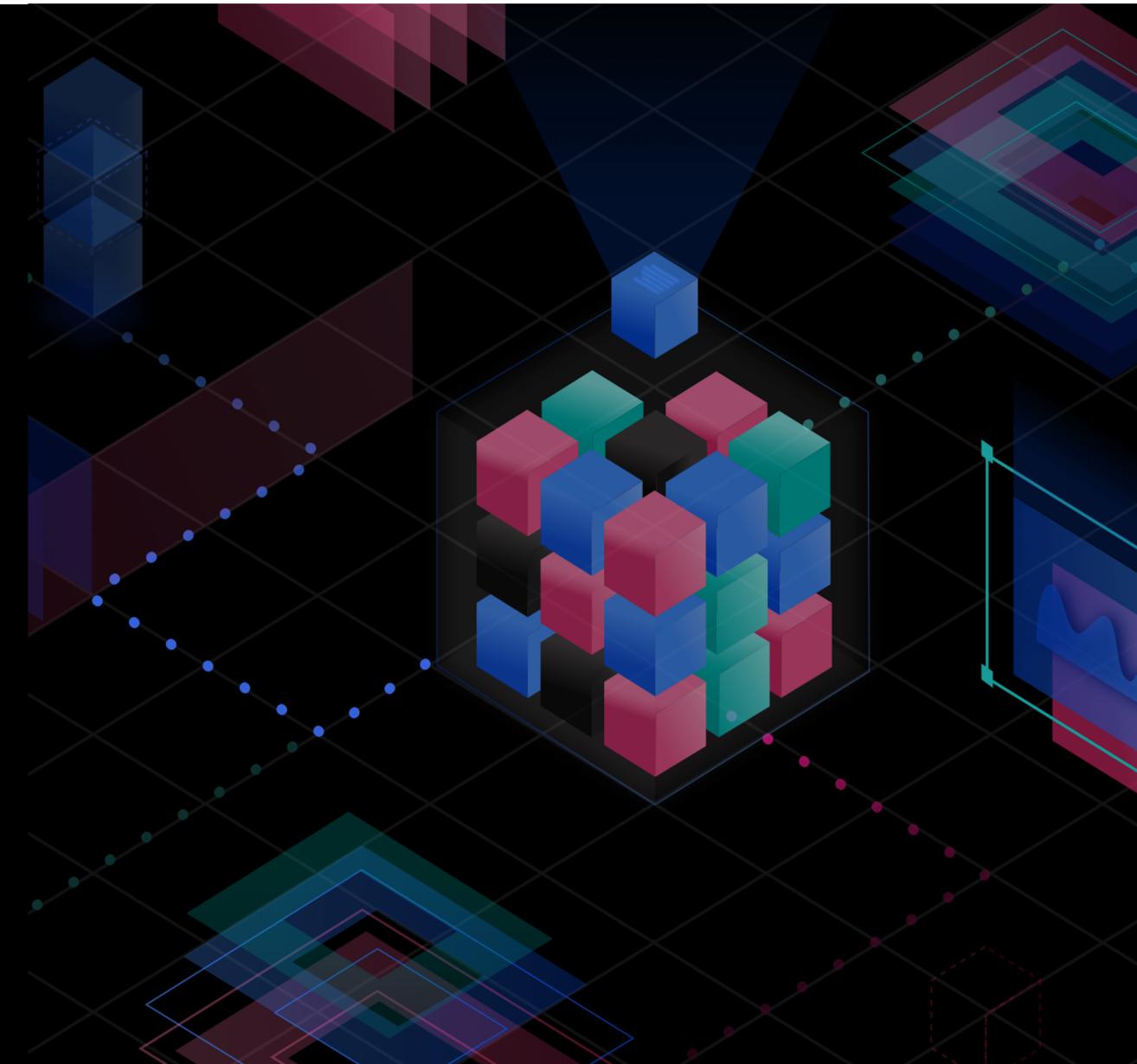


# Analyze Data

---

Rachana Vishwanathula  
Hybrid Cloud Build Team  
[rachvis1@in.ibm.com](mailto:rachvis1@in.ibm.com)



# The AI Ladder

## IBM's prescriptive approach to the journey to AI

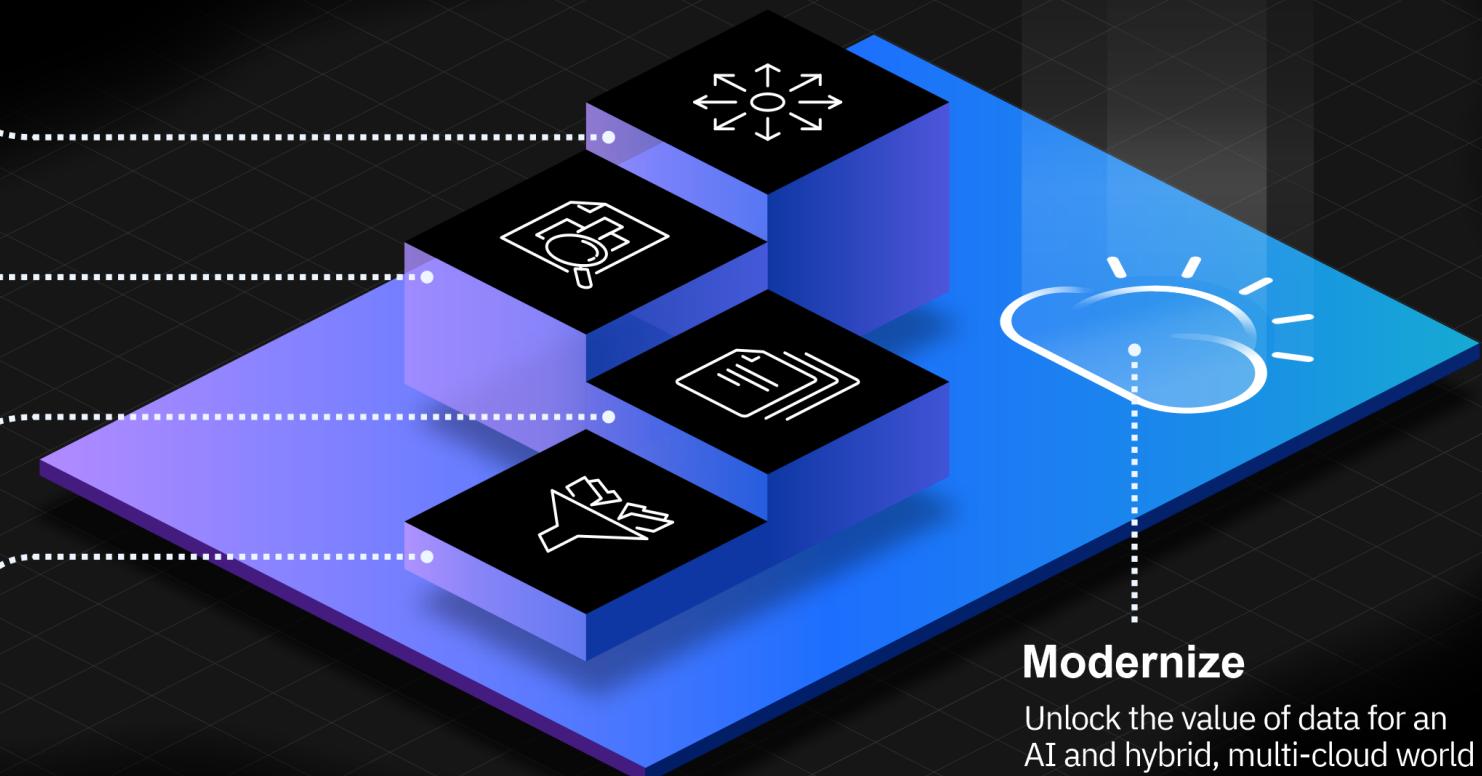
**Infuse**  
Operationalize AI  
throughout the enterprise

**Analyze**  
Build and scale AI with  
trust and transparency

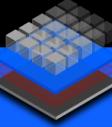
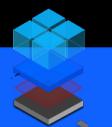
**Organize**  
Create a business-ready  
analytics foundation

**Collect**  
Make all data simple  
and accessible

**Modernize**  
Unlock the value of data for an  
AI and hybrid, multi-cloud world



# Cloud Pak for Data v4.0 Packaging

		AI Ladder Rung	Base Services	IBM Cartridges
	Collect			IBM Cartridges
Collect		Db2 Warehouse Data Virtualization Db2 BigSQL Netezza Performance Server IBM Streams Analytics Engine for Apache Spark Hadoop Execution Engine		Db2 AE/SE Informix
Organize		Watson Knowledge Catalog (including IGC) Information Analyzer (included in WKC) Data Management Console Data Privacy (Beta) IBM Match360 with Watson Guardium (Integration)		Master Data Management Product Master DataStage Information Server Knowledge Accelerators
Analyze		Watson Studio (includes Data Refinery) Watson Machine Learning (includes AutoAI) Watson Machine Learning Accelerator Watson OpenScale SPSS Modeler Decision Optimization		-
Infuse	Cognos Dashboards Embedded			Cognos Analytics Planning Analytics with Watson Watson Assistant Watson Discovery Watson Speech Services Financial Crimes Insights OpenPages with Watson Open Data for Industries Financial Services Workbench

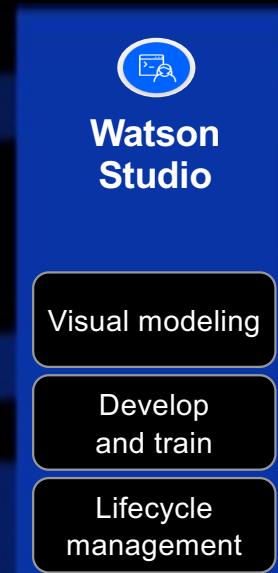
# Analyze – Data science

## Operationalizing AI with trust & transparency

Prepare and Organize Data



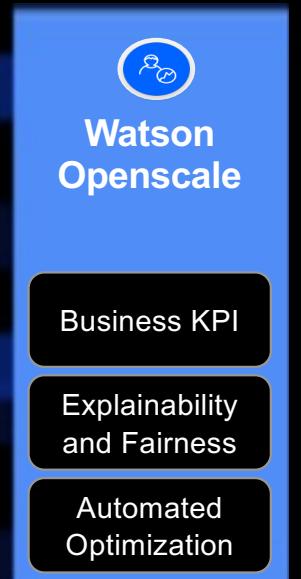
Build and Train AI Models



Deploy and Run AI Models



Manage and Operate Trusted AI

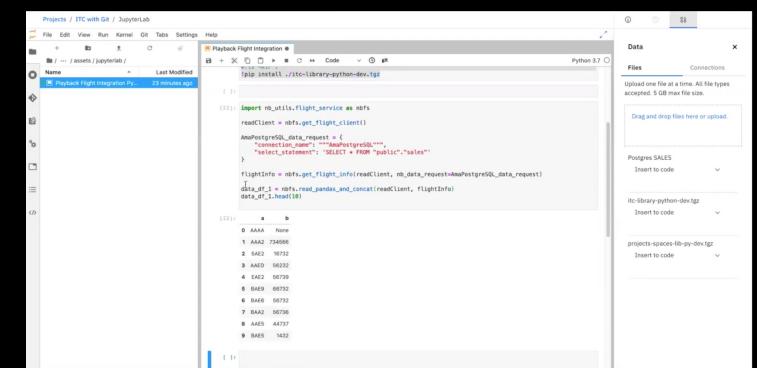
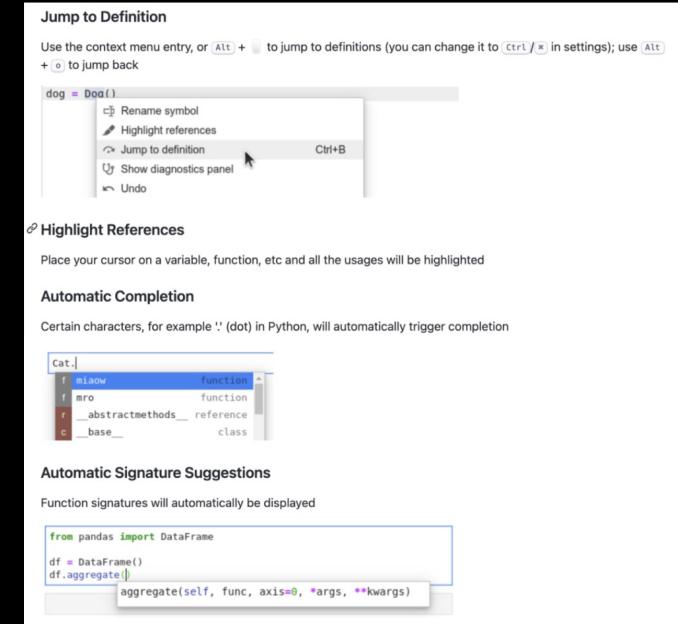


Automated AI Lifecycle

# Watson Studio in Cloud Pak for Data 4.0

## Notebooks and Runtimes

- JupyterLab and Jupyter Notebooks
  - Support for JupyterLab 3.0 version.
  - With Elyra 2.0 updates Jupyter Lab integrates with Language server protocol to support autocomplete, code navigation, hover suggestion, code linting
- Connection support in Notebook
  - Notebook insert to code feature can support all connections on the platform via Flight service integration.
  - This will generate cleaner code, no credentials exposed in the code and performance improvements
- Spark related enhancements
  - Ability to select more than 10 Executors in UI.
  - User can launch Spark UI from Jupyter notebooks to monitor Spark application.
  - Users can see progress of cell executions in Spark Notebooks.
- R Studio
  - Users can use Custom image with R Studio.



# Watson Studio in Cloud Pak for Data 4.0

## Data Refinery and Jobs

- Data Refinery
  - Improved scalability through reduced vCPU per user
  - Spark 3.0 support
  - Improved Interactive and Job performance for Data refinery flows
- Jobs
  - Jobs users can define retention policy for job runs and logs .
  - Users can take an action on multiple assets, monitor Job run status & average duration for active runs on the operations view.

31 items selected					
<input checked="" type="checkbox"/> Start time	Status	In progress	Job Project	Asset type	
Apr 12, 2021 4:36:17 PM Started by Andy Lac	<span>Paused</span>	00:00:28 00:00:22 x	DR_status_reasons status-reasons	Data Refinery Flow	<a href="#">Cancel runs</a> <a href="#">Cancel</a>
Apr 05, 2021 11:46:31 AM Started by Andy Lac	<span>Starting</span>	217:59:09	testing_status_reasons status-reasons	Notebook	
Apr 01, 2021 2:30:41 PM Started by Scheduler	<span>Starting</span>	311:15:00 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:29:28 PM Started by Scheduler	<span>Starting</span>	311:16:18 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:28:16 PM Started by Scheduler	<span>Starting</span>	311:17:25 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:27:33 PM Started by Scheduler	<span>Starting</span>	311:18:08 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:26:56 PM Started by Scheduler	<span>Starting</span>	311:18:45 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:25:40 PM Started by Scheduler	<span>Starting</span>	311:20:01 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:24:25 PM Started by Scheduler	<span>Starting</span>	311:21:16 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	
Apr 01, 2021 2:23:11 PM Started by Scheduler	<span>Starting</span>	311:22:30 00:01:33 x	van-test-cancel-nb unchanged-bss-id-PR	Notebook	<a href="#">Comment</a>

## Asset Browser

- Support for search, filter, sort, multi select.

# Project and Deployment Space in Cloud Pak for Data 4.0

## User Management

- User Groups:
  - Create a user group and assign set of roles and permission to multiple users.
- Create permission
  - Administrator can choose to restrict users from creating new projects, spaces
- Manage All permission
  - Enable user to see a list of all deployment spaces/projects and view deployment activity and active runtimes across all spaces and projects respectively.
  - Users with this permission can join any project as an administrator so that they can delete unused projects and ensure that active projects have at least one owner.
- Monitor Permission
  - Enable users to see all active jobs and deployments across all spaces and active runtimes for all projects from the Active runtimes page

The image contains two screenshots of the Cloud Pak for Data 4.0 User Management interface. The top screenshot shows the 'New user group' creation screen, where the 'Details' tab is selected. It includes fields for 'Name' (set to 'Users Group 1') and 'Description (optional)'. The bottom screenshot shows the 'Add permissions' screen, where the 'Monitor project workloads' checkbox under the 'Deployments' category is selected. Both screenshots show standard UI elements like 'Cancel', 'Back', and 'Next' buttons.

# Watson Studio in Cloud Pak for Data 4.0

## Train, Deploy and Trust

- Model versioning
  - Surface model versions in the UI
  - Edit deployment to refer another model version
- Deployment
  - Support for more databases in SPSS batch scoring:
    - Oracle (with SQL Pushback)
    - SnowFlake (with SQL Pushback)
    - PostgreSQL (with SQL Pushback)
    - Informix
    - SAP HANA
  - Notebooks Jobs in deployment spaces - Output of each job run saved independently
  - Support for Notifications on deployments
- Watson Studio – Monitor & Trust
  - Support for Explainability in Batch Environments
  - Bias Detection in Batch Environments
  - Support for Non-Hive Data Sources in Batch Environments

The top screenshot shows the 'Create a deployment' dialog in the IBM Cloud Pak for Data interface. It allows selecting an associated asset (Another Drug PMML), choosing a model version (Current, Version 3, or Version 4), and providing a deployment name and description.

The bottom screenshot shows the 'Create a job' dialog for a notebook named 'Sumits Notebook'. It includes tabs for 'Define details', 'Configure', 'Schedule', and 'Review and create'. Under 'Review and create', details like 'Associated asset: Sumits Notebook', 'Name: First Notebook Job in a Deployment Space', and 'Environment: Default Python 3.7' are shown. The 'Schedule' tab shows a repeating job starting on 06/17/2021 at 3:20 PM, repeating every 3 minutes until 06/17/2021 at 3:34 PM.

# Watson Studio in Cloud Pak for Data 4.0

## SPSS Modeler Flows

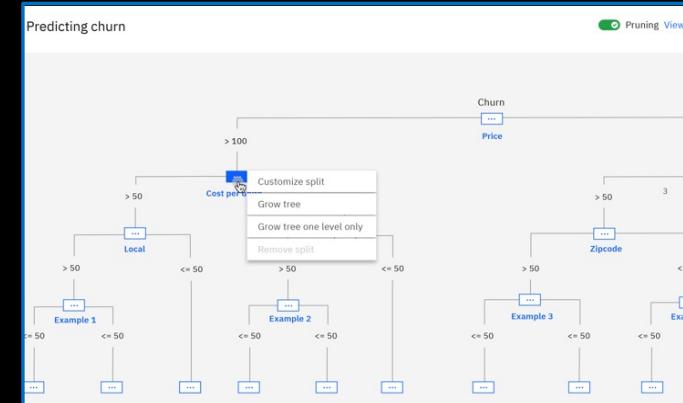
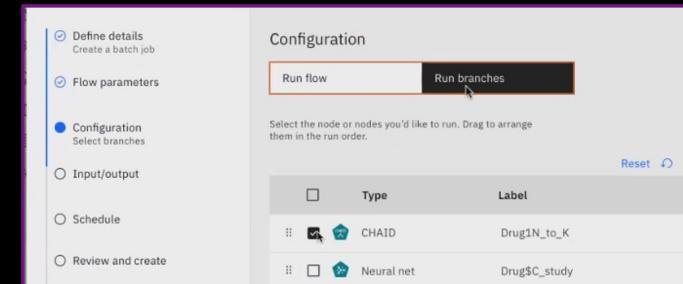
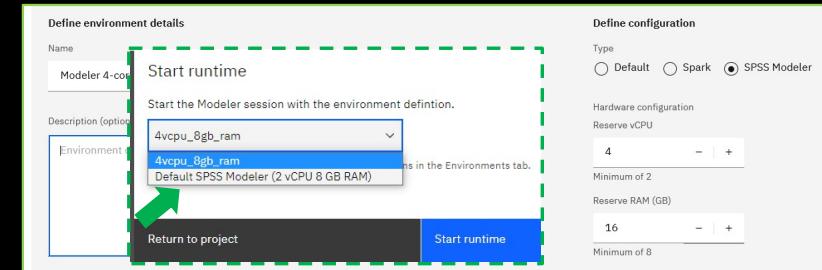
- New support for **custom runtime environment sizes**
  - **Note:** Default SPSS Modeler environment size reduced to 2 vCPU & 8gb RAM *from* 4 vCPU & 12gb RAM

*Functional improvements:*

- Jobs enhancements: **custom environments support**
- Jobs enhancements: option to **run specific branches**
- Support for **R 4.0**; ability to **execute Python & R** within a Modeler Stream
- Support for **Interactive Decision Tree**; Sim Eval node; non-negative Regression
- New Continuous ML capability

*Usability Improvements:*

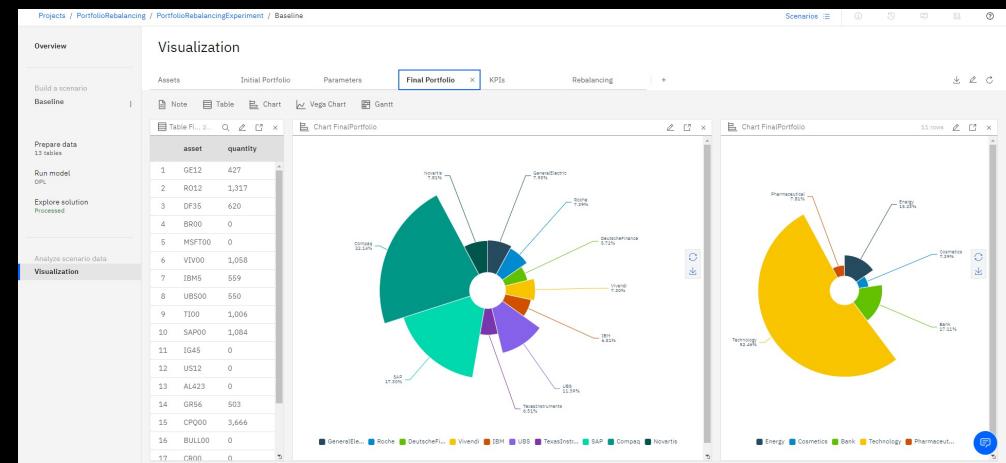
- **Undo & Redo; Copy & Paste** nodes across streams\*
- Analysis node + Statistics node **output presentation improvements**



# Watson Studio in Cloud Pak for Data 4.0

## Decision Optimization and Hadoop Execution Engine

- Decision Optimization
  - Support for C# models – early access
  - Solve from DO Experiment UI is using CPLEX 20.1
  - CPLEX 20.1 available in WML
  - Audit logging and security
  - Accessibility
- Hadoop Execution Engine
  - HEE users can securely connect to Cloudera Data Platform 7.1.5 for data access and offloading workloads.
  - Support insert to code for HEE connectors
  - JupyterLab working with HEE connectors
  - Notebook running in spaces with HEE connectors



# Watson Studio in Cloud Pak for Data 4.0

## Elastic GPU Computing

- Deep Learning libraries updates including:
  - NVIDIA CUDA Toolkit 11.0 that supports for NVIDIA A100 GPU
- High Availability with multiple active replicas
- Training Backup and Restore
- Auditing (Audit Logging)
- Openshift Container Storage 4.6.0
- Enhanced security with TLS 1.2.0

ID & Name	Status	Type	CPU	GPU	Pack ID	User	Submit
bjwjliao4-31 pytorch-mnist-edt-gpu	Pending	Training	- / 0	- / 4	-	wmla-user	10/20/2020, 03:06 PM
bjwjliao4-30 pytorch-mnist-edt-gpu	Running	Training	0.10 / 0.1	4 / 4	-	wmla-user	10/20/2020, 03:05 PM

ID & Name	Status	Type	CPU	GPU	Pack ID	User	Submit
bjwjliao4-31 pytorch-mnist-edt-gpu	Running	Training	0.10 / 0.1	2 / 4	-	wmla-user	10/20/2020, 03:06 PM
bjwjliao4-30 pytorch-mnist-edt-gpu	Running	Training	0.10 / 0.1	2 / 4	-	wmla-user	10/20/2020, 03:05 PM

ID & Name	Status	Type	CPU	GPU	Pack ID	User	Submit
bjwjliao4-31 pytorch-mnist-edt-gpu	Running	Training	0.10 / 0.1	3 / 4	-	wmla-user	10/20/2020, 03:06 PM
bjwjliao4-30 pytorch-mnist-edt-gpu	Running	Training	0.10 / 0.1	1 / 4	-	wmla-user	10/20/2020, 03:05 PM

# Watson Studio in Cloud Pak for Data 4.0

## AutoAI

- Code generation GA
- Python client GA
- Support free text columns in tabular data
- Database support for data ingestion
  - Db2
  - MySQL
  - MS SQL Server
  - PostgreSQL
  - Netezza

The screenshot shows a Jupyter Notebook interface titled "Pipeline 8 Notebook - AutoAI Notebook v1.15.0". The top bar includes the AutoAI logo and the text "Part of IBM Watson® Studio" and "Pipeline Notebook". The main content area has the following sections:

- Pipeline 8 Notebook - AutoAI Notebook v1.15.0**
- Consider these tips for working with an auto-generated notebook:**
  - Notebook code generated using AutoAI will execute successfully. If you modify the notebook, we cannot guarantee it will run successfully.
  - This pipeline is optimized for the original data set. The pipeline might fail or produce sub-optimum results if used with different data. If you want to use a different data set, consider retraining the AutoAI experiment to generate a new pipeline. For more information, see [Cloud Platform](#)
  - Before modifying the pipeline or trying to re-fit the pipeline, consider that the code converts dataframes to numpy arrays before fitting the pipeline (a current restriction of the preprocessor pipeline).
- Notebook content**

This notebook contains a Scikit-learn representation of AutoAI pipeline. This notebook introduces commands for getting data, training the model, and testing the model.
- Notebook goals**
  - Scikit-learn pipeline definition
  - Pipeline training
  - Pipeline evaluation
- Contents**

This notebook contains the following parts:

  - Setup**
    - [Package installation](#)
    - [AutoAI experiment metadata](#)
  - Pipeline inspection**
    - [Read training data](#)
    - [Train and test data split](#)
    - [Make pipeline](#)
    - [Train pipeline model](#)
    - [Test pipeline model](#)
  - Next steps**
    - [Copyrights](#)

# Beta Features In Watson Studio on Cloud Pak for Data 4.0

## ■ AutoAI time series forecasting

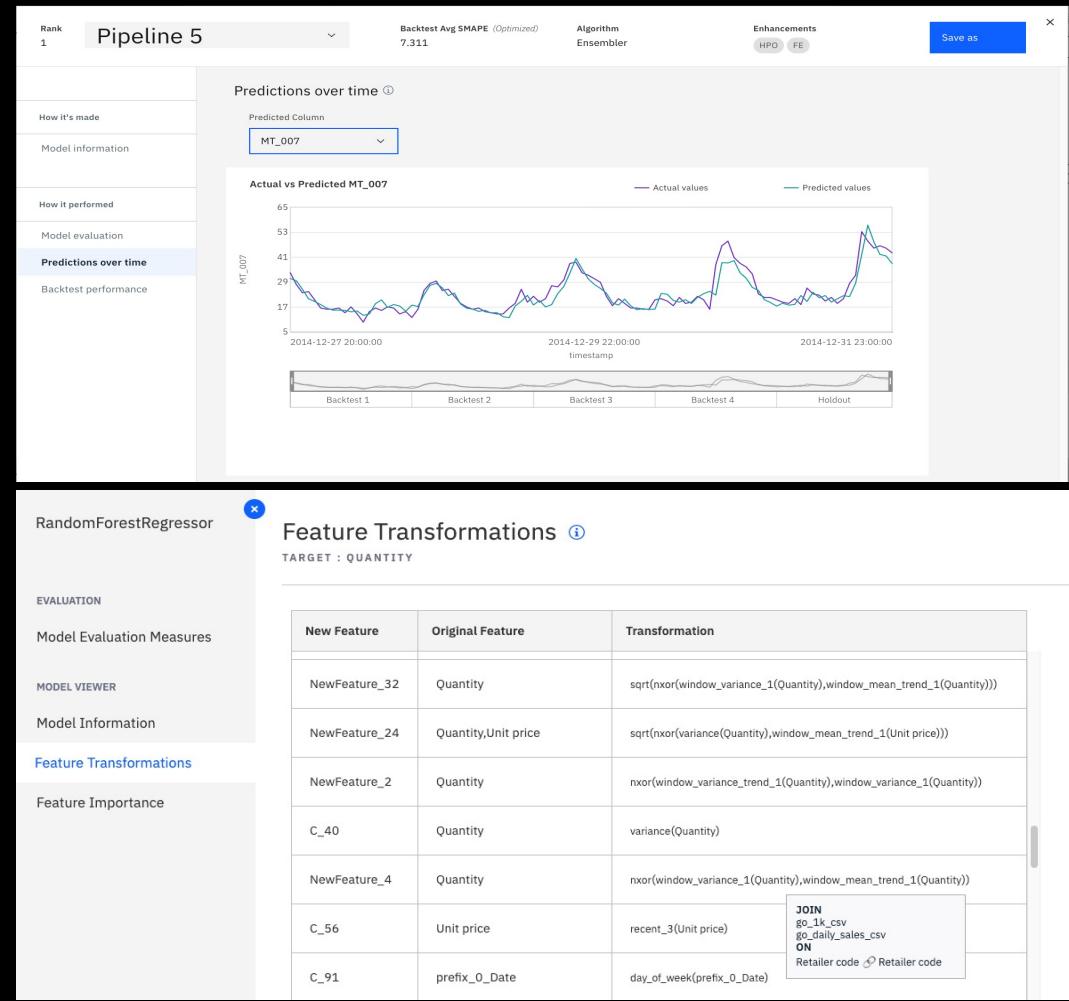
- Auto generate pipelines of up to 7 different time series algorithms
- Support multi-step forecasting
- Configurable back testing setup

## ■ AutoAI feature engineering on relational datasets

- Generate new features with original features from different datasets
- Drag and drop UI for configuring join relationships between multiple datasets
- Distribute workload with Spark

## ■ Federated Learning

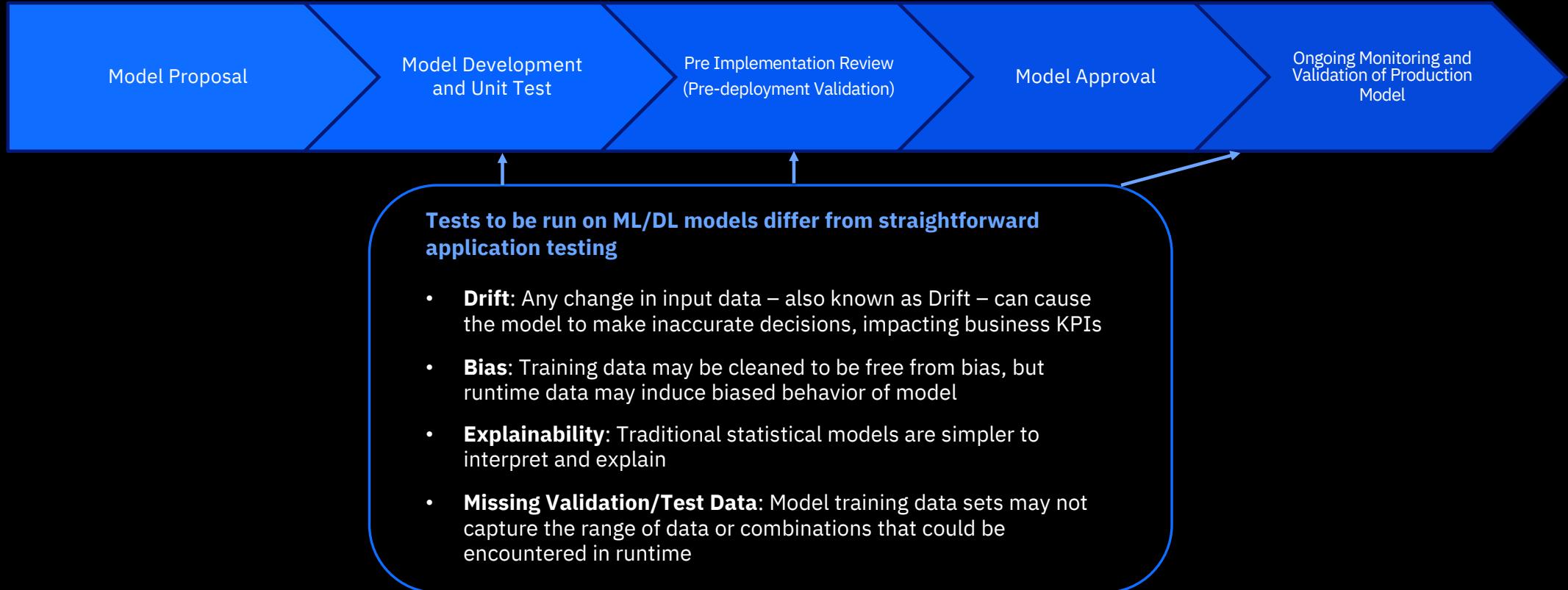
- PyTorch 1.7 & TensorFlow 2.4
- Early termination
- Quorum management
- PFNM support for PyTorch



# Model Development Cycle

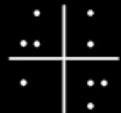
## *Challenges faced with ML/DL Models*

**Complexity and Assessment:** Lack of knowledge of methods used by Model Developers / Vendors along with inconsistent documentation and increased volume of model change.



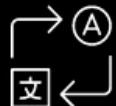
# Watson OpenScale

Validate and monitor AI models, deployed anywhere, to help comply with regulations, address internal safeguards, and mitigate business risk



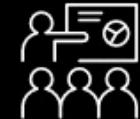
## Production monitoring for compliance and safeguards

- Mitigate biased model behavior
- Explain model decisions
- Validate and control risk



## Ensure that models are resilient to changing situations

- Detect drift during runtime
- Generate specific model retraining inputs



## Align model performance with business outcomes

- Actionable metrics and alerts

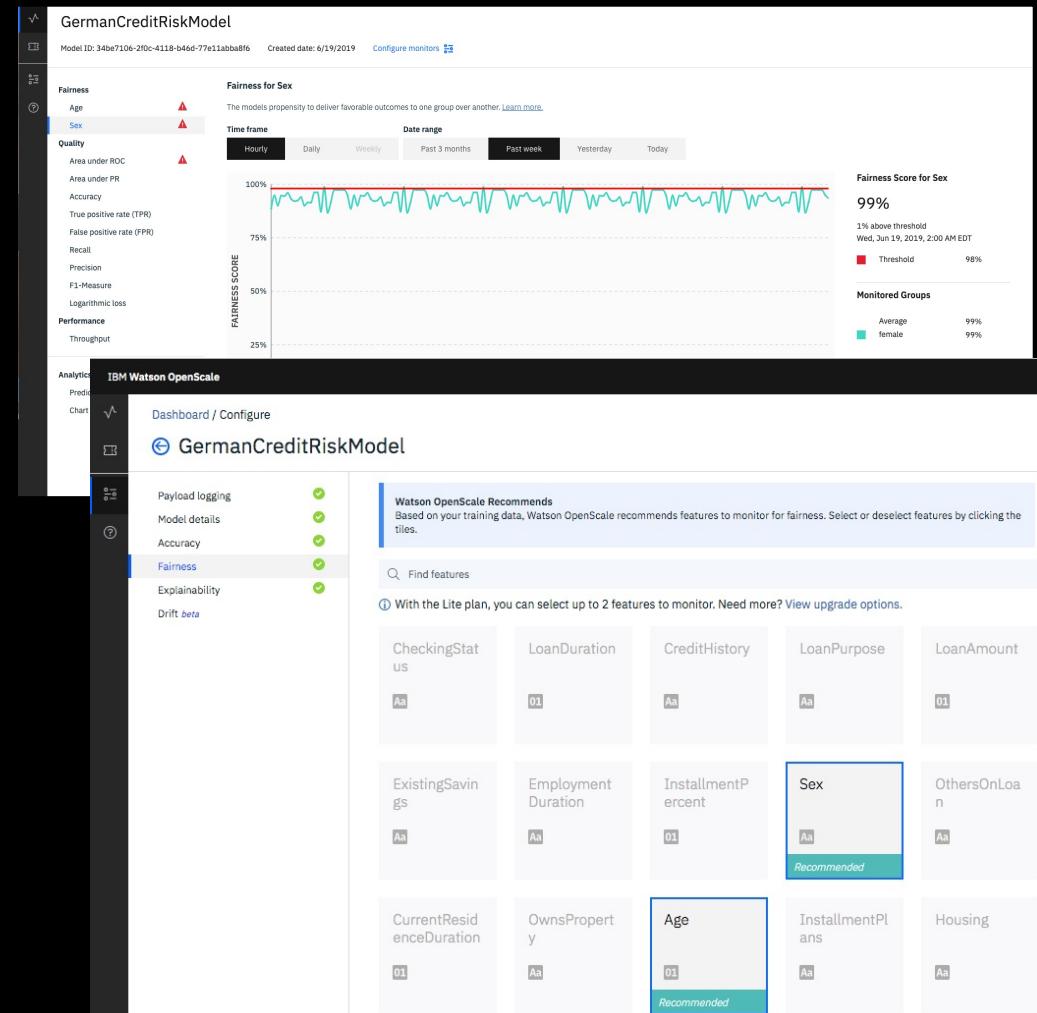
# Bias Detection

OpenScale enables enterprises to enforce fairness in their model's outcome by analyzing transactions in production and finding biased behavior by the model

It pinpoints the source of bias and actively mitigates the biases found in production environment

## Value:

- Automatically recommend common protected attributes to monitor during production
- Detect biases in runtime in order to catch impacts on business applications and compliance requirements without time consuming, manual data analysis
- Metrics and data to help data scientists further troubleshoot issues in data sets or models
- Mitigate biases in runtime in order to enforce regulatory or enterprise fairness guardrails in real time



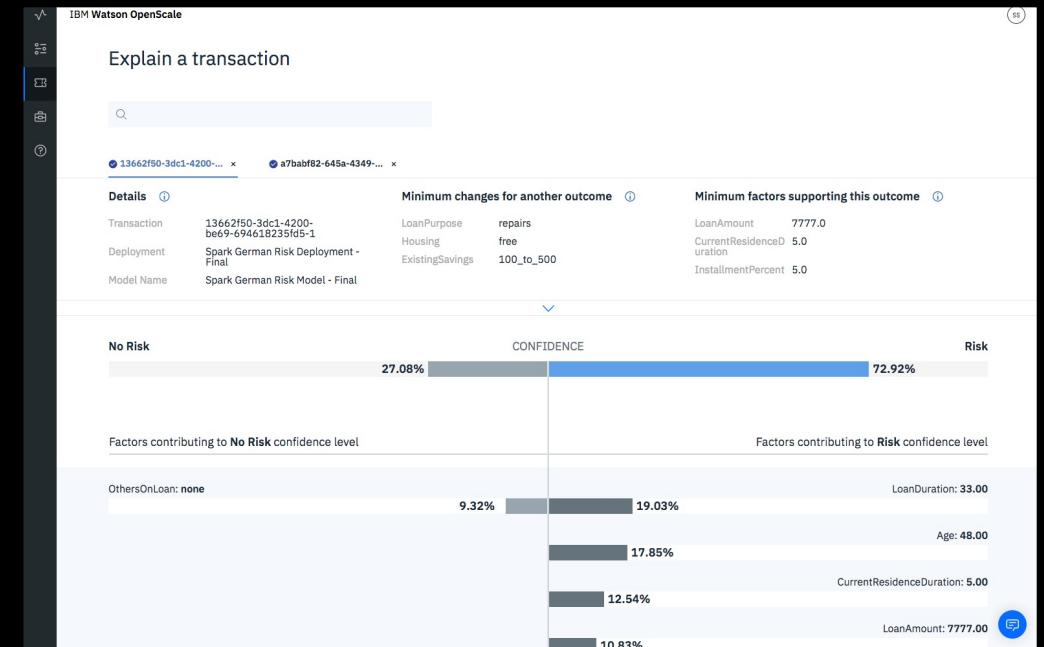
# Explainability

OpenScale records every individual transaction and drills down into its working to explain how the model makes decisions

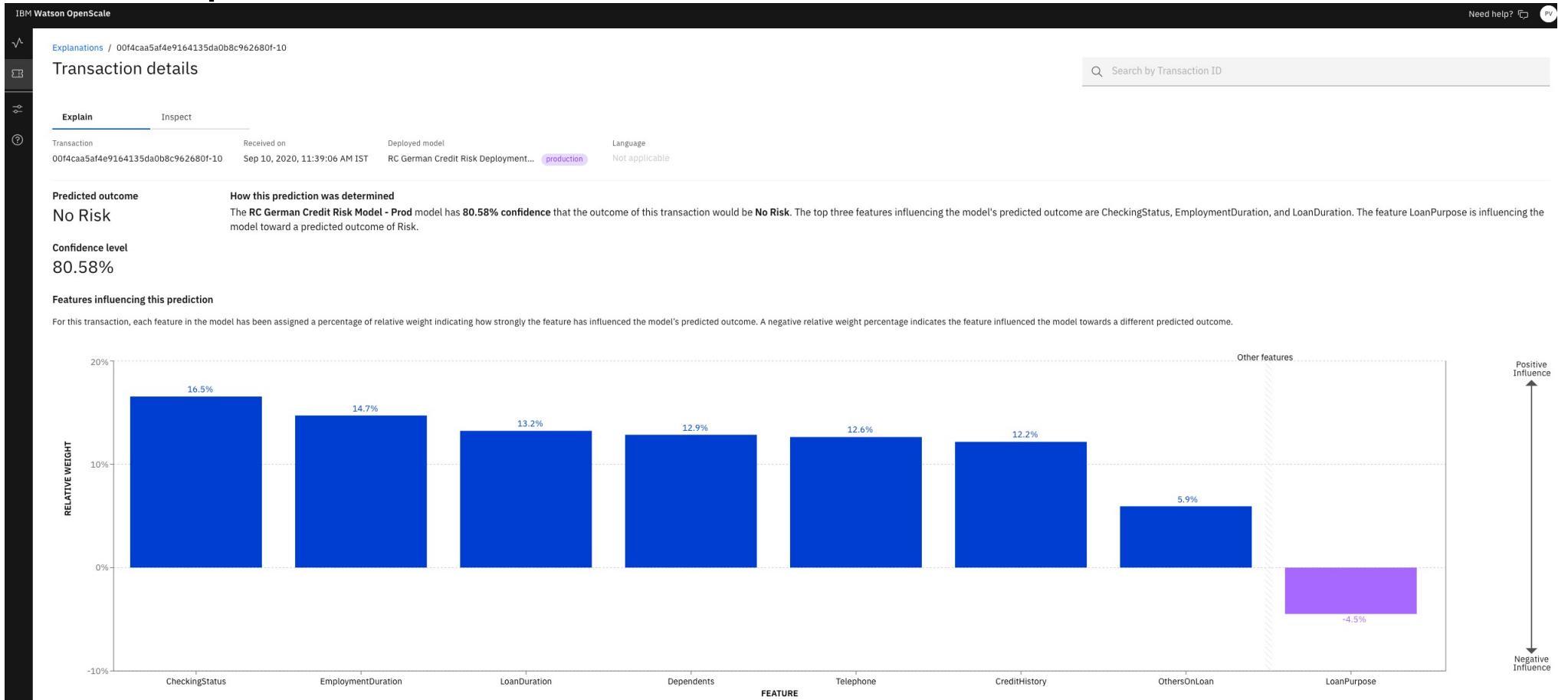
It provides a simple explanation that is user friendly and interactive

## Value:

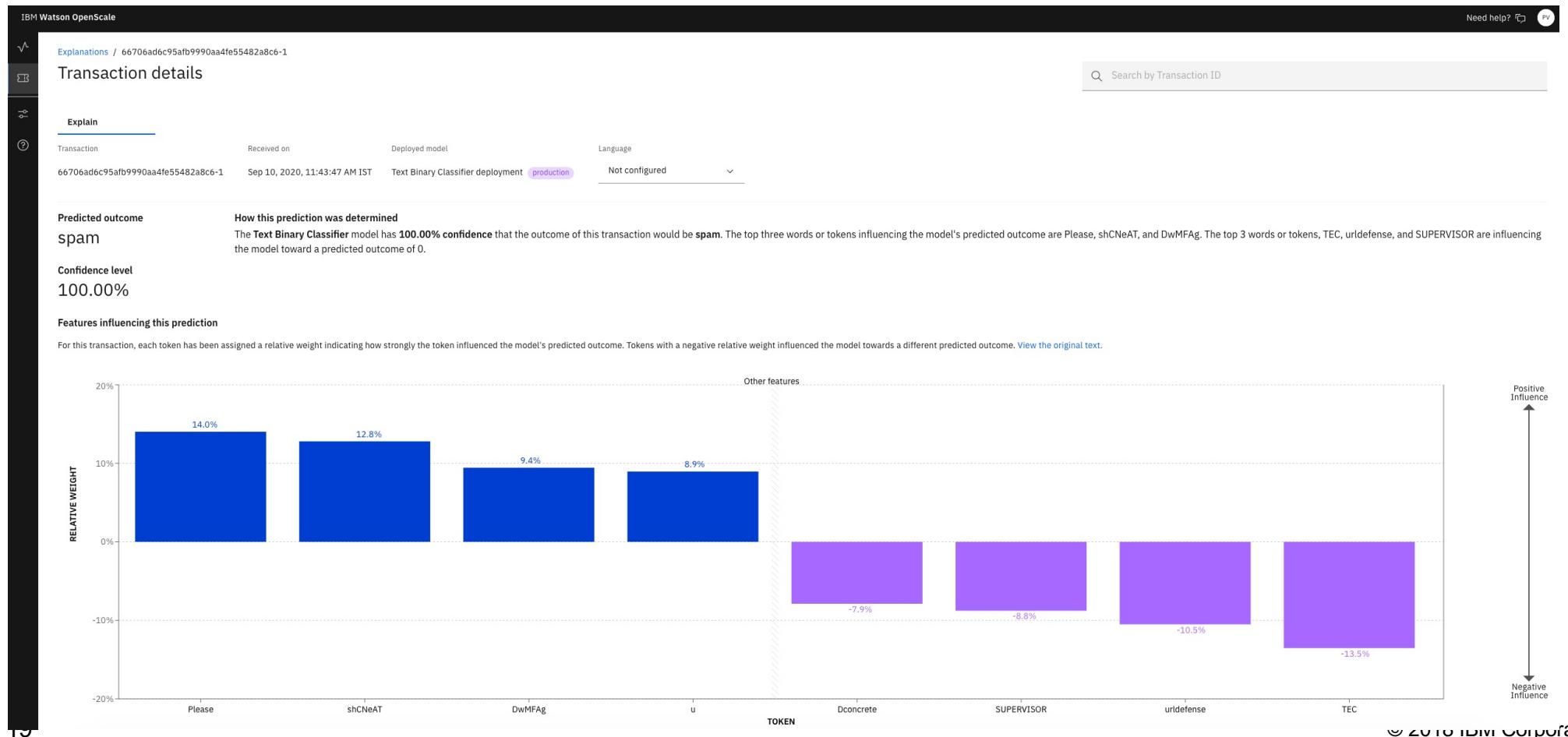
- Explain individual transaction level decisions made by the model in run time, including details about most important attributes and their values in order to assist in compliance and customer care situations
- Analyze individual transactions in a what-if manner in order to understand how model behavior will change in different business situations



# Lime Explanation for structured data



# Explanation for unstructured text



# Explanation for unstructured image

IBM Watson OpenScale      Need help?  

[Explanations / 29c1e2acd1fb469ea23c5ca247a72f16-1](#)

**Transaction details**

**Explain**

Transaction      Received on      Deployed model      Language  
 29c1e2acd1fb469ea23c5ca247a72f16-1      Sep 10, 2020, 12:20:52 PM IST      Fashion MNIST Model Deployment - production      Not applicable

**Predicted outcome**      **How this prediction was determined**  
 2      The **Fashion MNIST Model** model has **96.84% confidence** that the outcome of this transaction is 2. The top image regions influencing the model's predicted outcome are displayed here.

**Confidence level**  
**96.84%**

**Features influencing this prediction**  
 For this transaction, the image was separated into segments. Each segment was altered to measure influence of the section on the model outcome. Sections that influenced the model's predicted outcome are presented as positive zones. Sections that influenced the model towards a different predicted outcome are presented as negative zones.



Origin Image



Positive zones



Negative zones

# Contrastive Explanation

IBM Watson OpenScale

Need help?

Explanations / 00f4caa5af4e9164135da0b8c962680f-10

Transaction details

[Search by Transaction ID](#)

Explain Inspect

Reaching a different predicted outcome

For the model to have predicted a different outcome for this transaction, the value of all listed features would need to change to the indicated minimum value. Note that changing a feature value by more than the minimum value may affect the minimum change of other features for the model to predict a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction.

Analyze controllable features only

Feature	Original value	New value	Value for a different outcome	Importance
Sex	male	male	female	1.00
CheckingStatus	no_checking	no_checking	no_checking	0.00
LoanDuration	25	25	25	0.00
CreditHistory	prior_payments_delayed	prior_payments_delayed	prior_payments_delayed	0.00
LoanPurpose	furniture	furniture	furniture	0.00
LoanAmount	7279	7279	7279	0.00
ExistingSavings	100_to_500	100_to_500	100_to_500	0.00
EmploymentDuration	4_to_7	4_to_7	4_to_7	0.00
InstallmentPercent	4	4	4	0.00
Predicted outcome <b>No Risk</b>	Confidence <b>80.58%</b>	Predicted outcome <b>No Risk</b>	Confidence <b>80.58%</b>	Predicted outcome <b>Risk</b>
				Confidence <b>73.81%</b>

# Thanks

---

Rachana Vishwanathula  
Hybrid Cloud Build Team  
[rachvis1@in.ibm.com](mailto:rachvis1@in.ibm.com)

