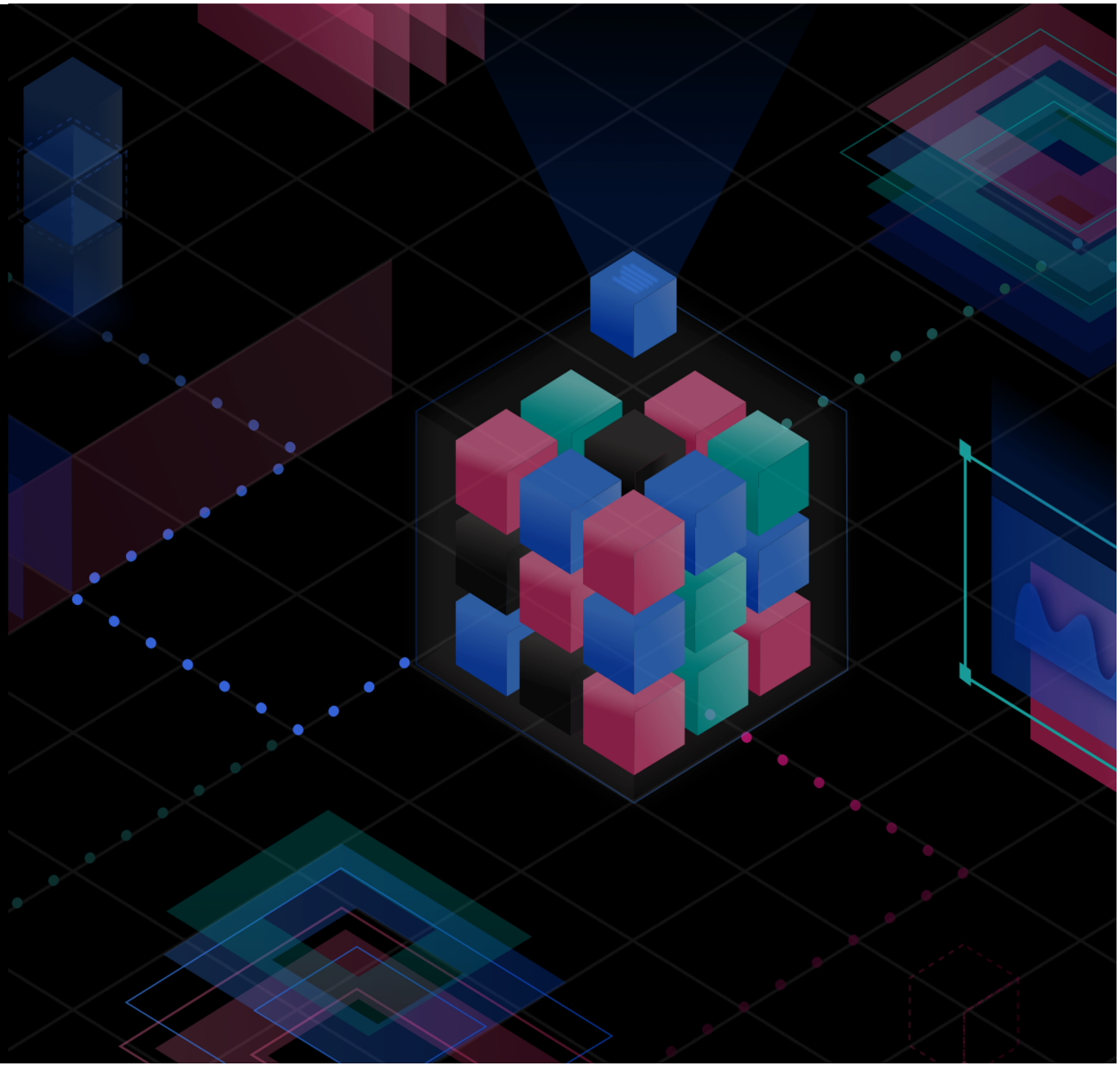
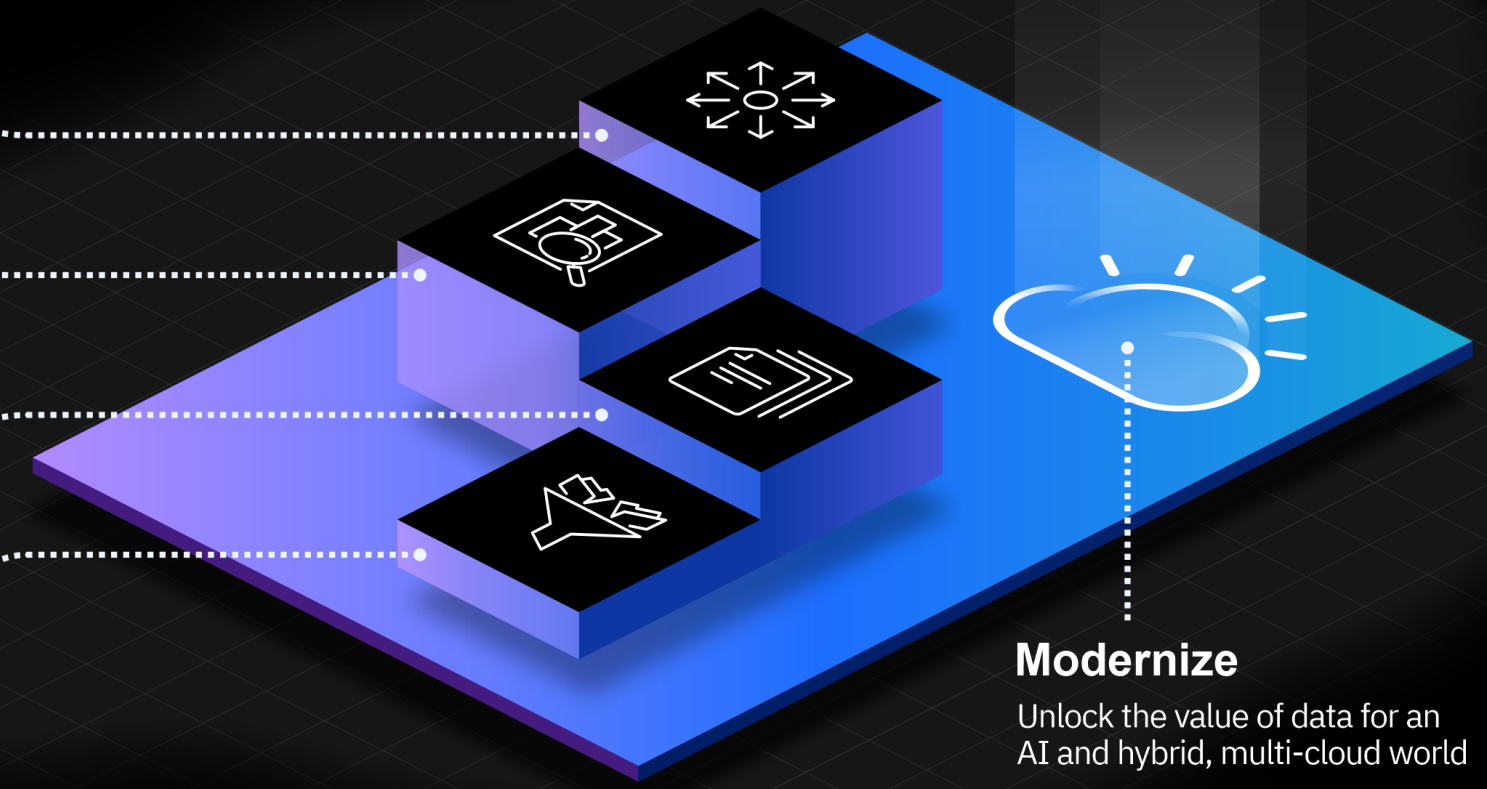
Analyze Data



Rachana Vishwanathula
Hybrid Cloud Build Team
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

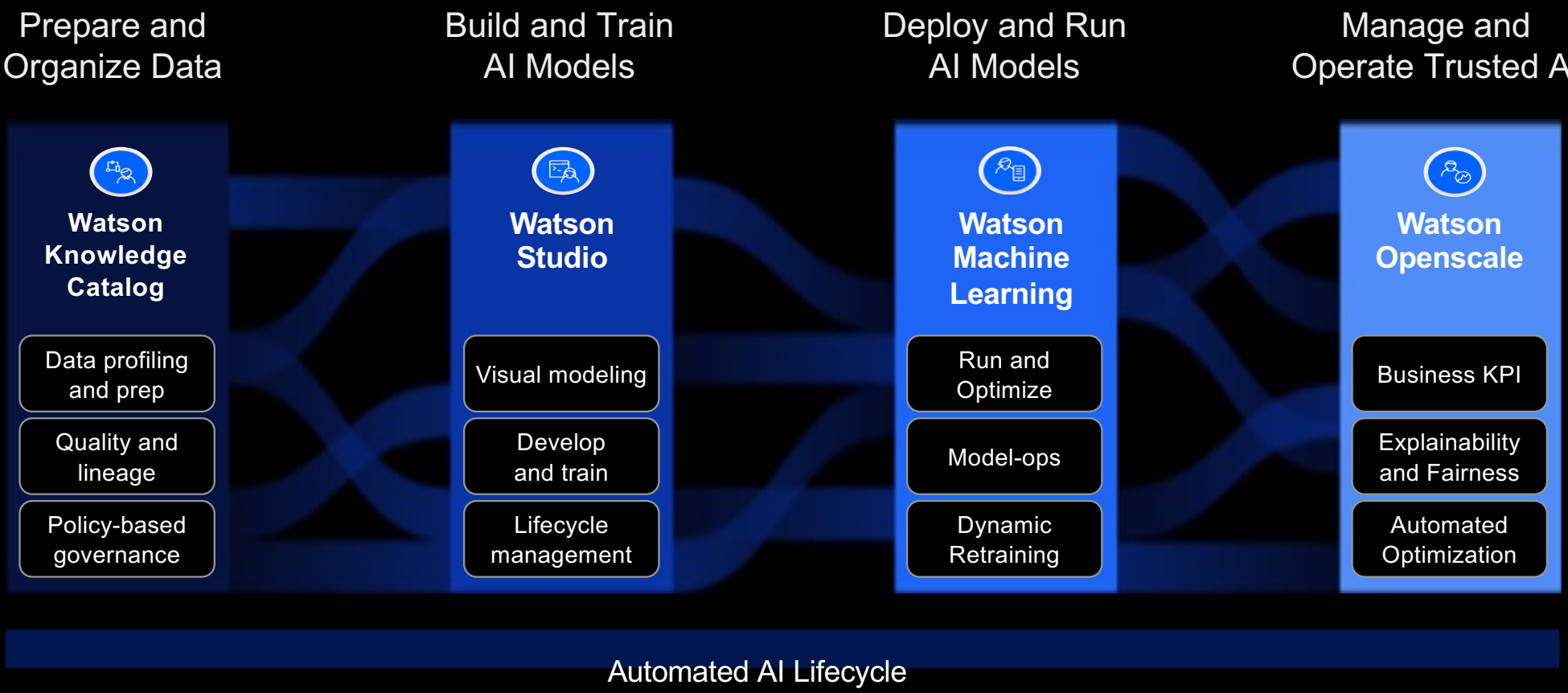
Modernize

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	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



Watson Studio in Cloud Pak for Data 4.0

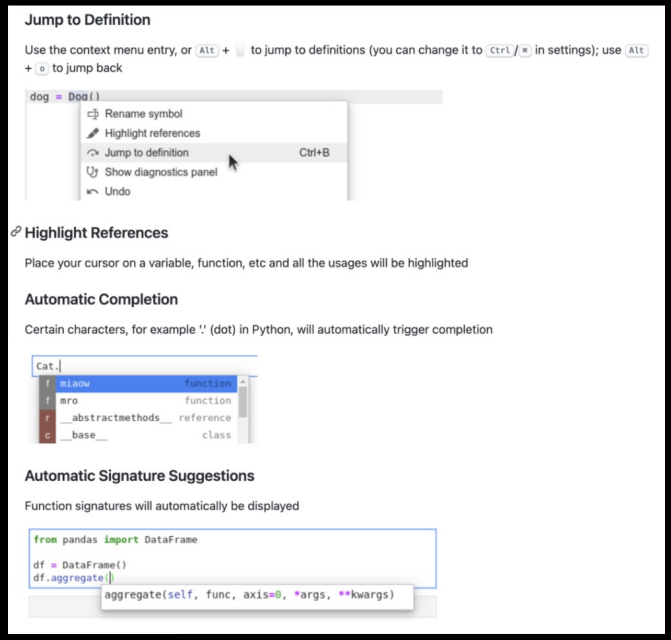
Notebooks and Runtimes

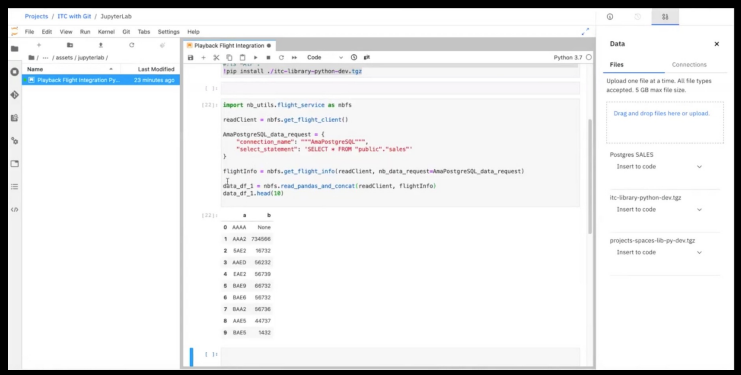
- JupyterLab and Jupyter Notebooks
  - Support for JupyterLab 3.0 version.
  - With Elyra 2.0 updates Jupyter Lab integrates with Language server protocol to support autocompletion, code navigation, hover suggestion, code linting

- 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.



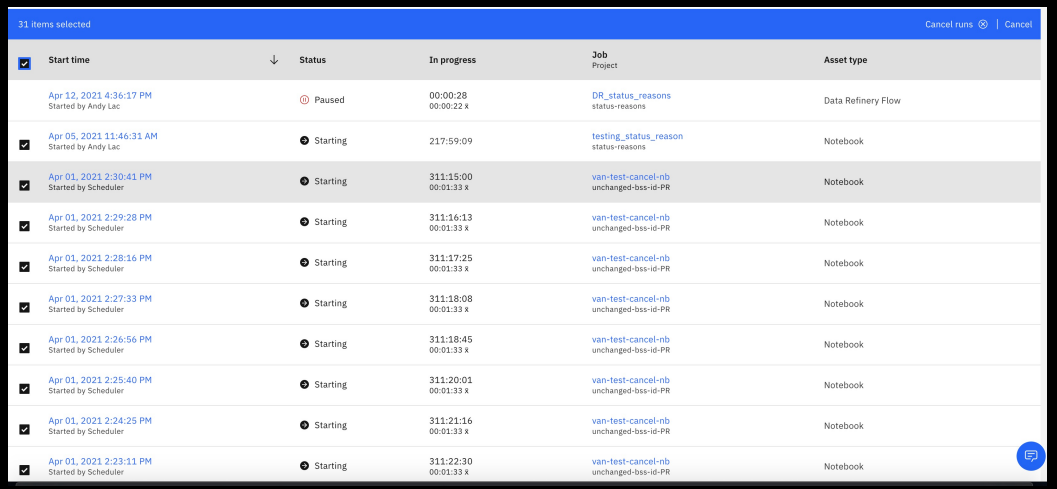


Watson Studio in Cloud Pak for Data 4.0

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.



Asset Browser

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

Project and Deployment Space in Cloud Pak for Data 4.0

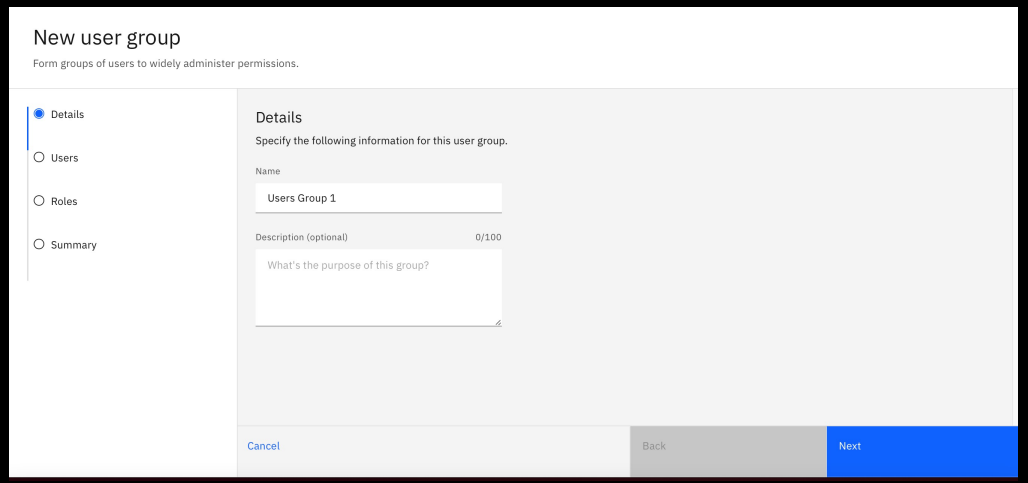
User Management

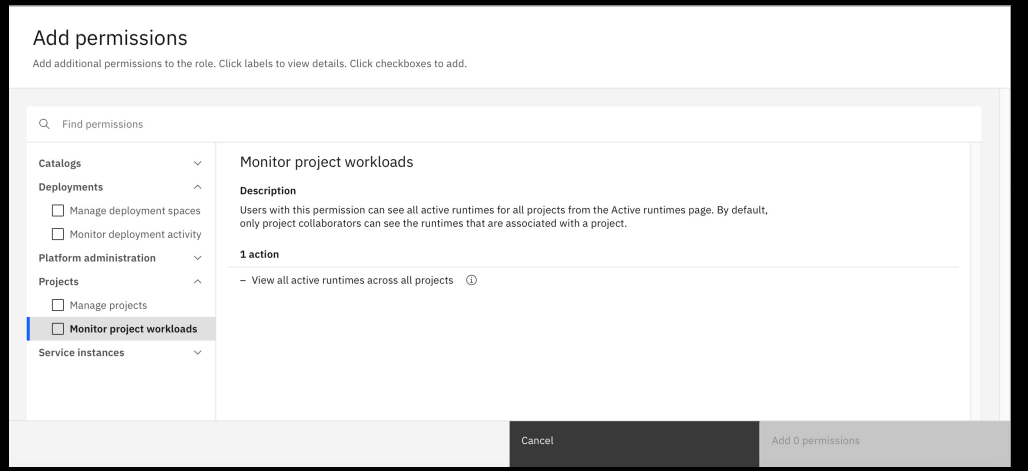
- User Groups:
  - Create a user group and assign set of roles and permission to multiple users.

- 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





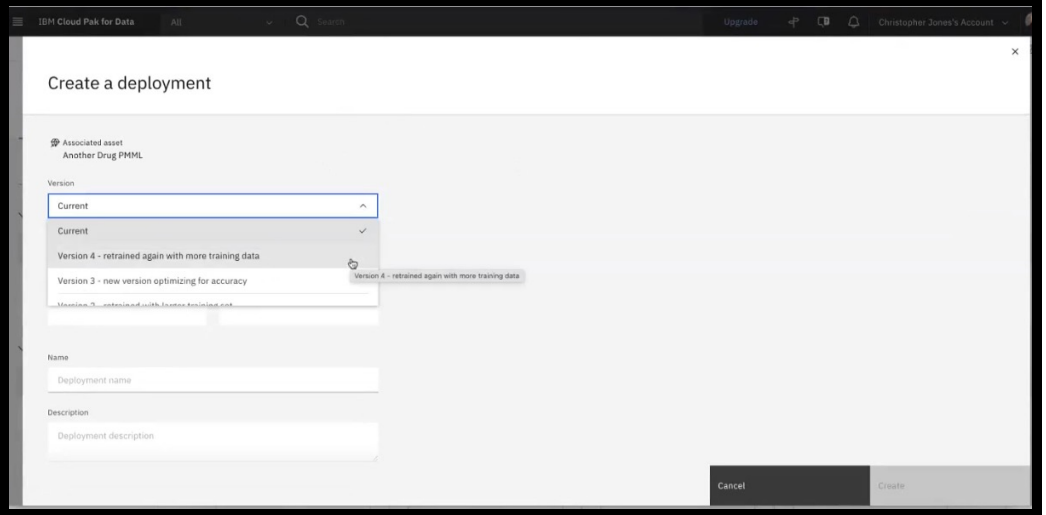
Watson Studio in Cloud Pak for Data 4.0

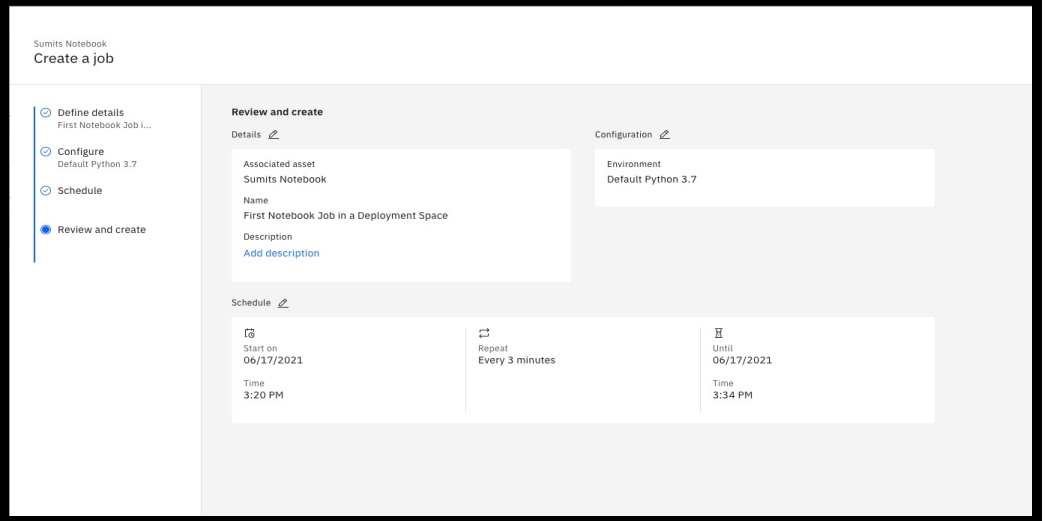
Train, Deploy and Trust

- 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





Watson Studio in Cloud Pak for Data 4.0

SPSS Modeler Flows

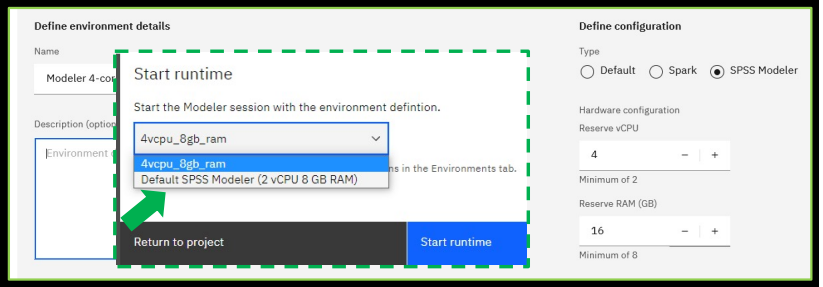
- New support for **custom runtime environment sizes**
  - ***Note***: *Default SPSS Modeler environment size reduced to 2 vCPU & 8gb RAM **from** 4 vCPU & 12gb RAM*

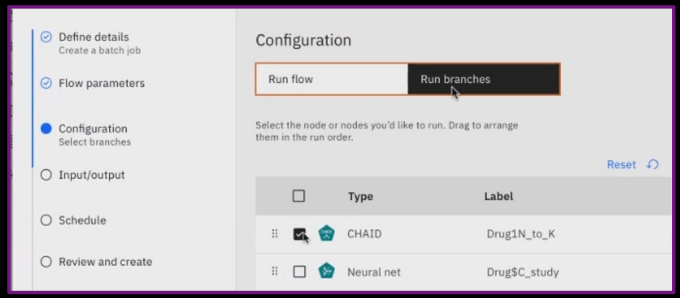
Functional improvements:

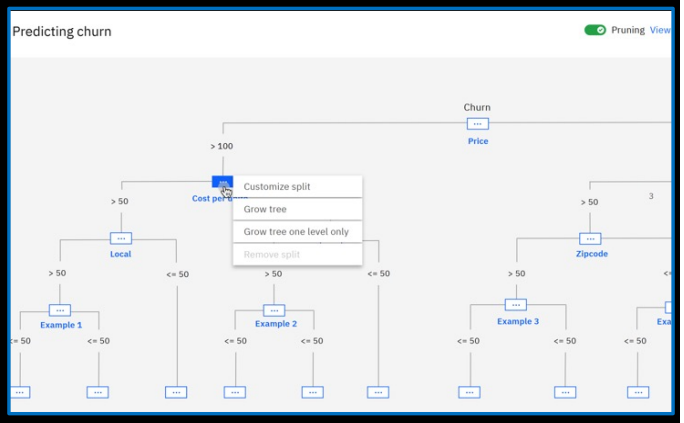
- Jobs enhancements: **custom environments support**
- Jobs enhancements: option to **run specific branches**
- Support for **R 4.0**; ability to **execute Python & R** within a Modeler Stream
- Support for **Interactive Decision Tree**; Sim Eval node; non-negative Regression
- New Continuous ML capability

Usability Improvements:

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





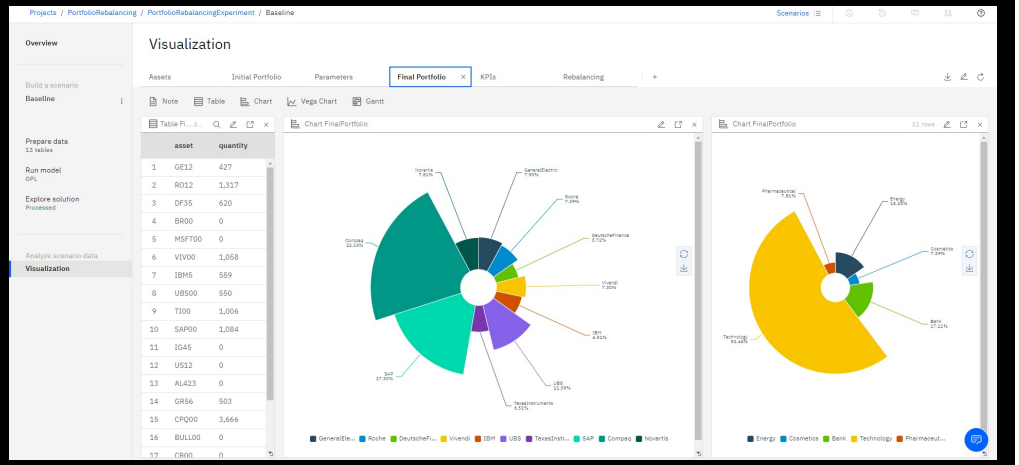


Watson Studio in Cloud Pak for Data 4.0

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

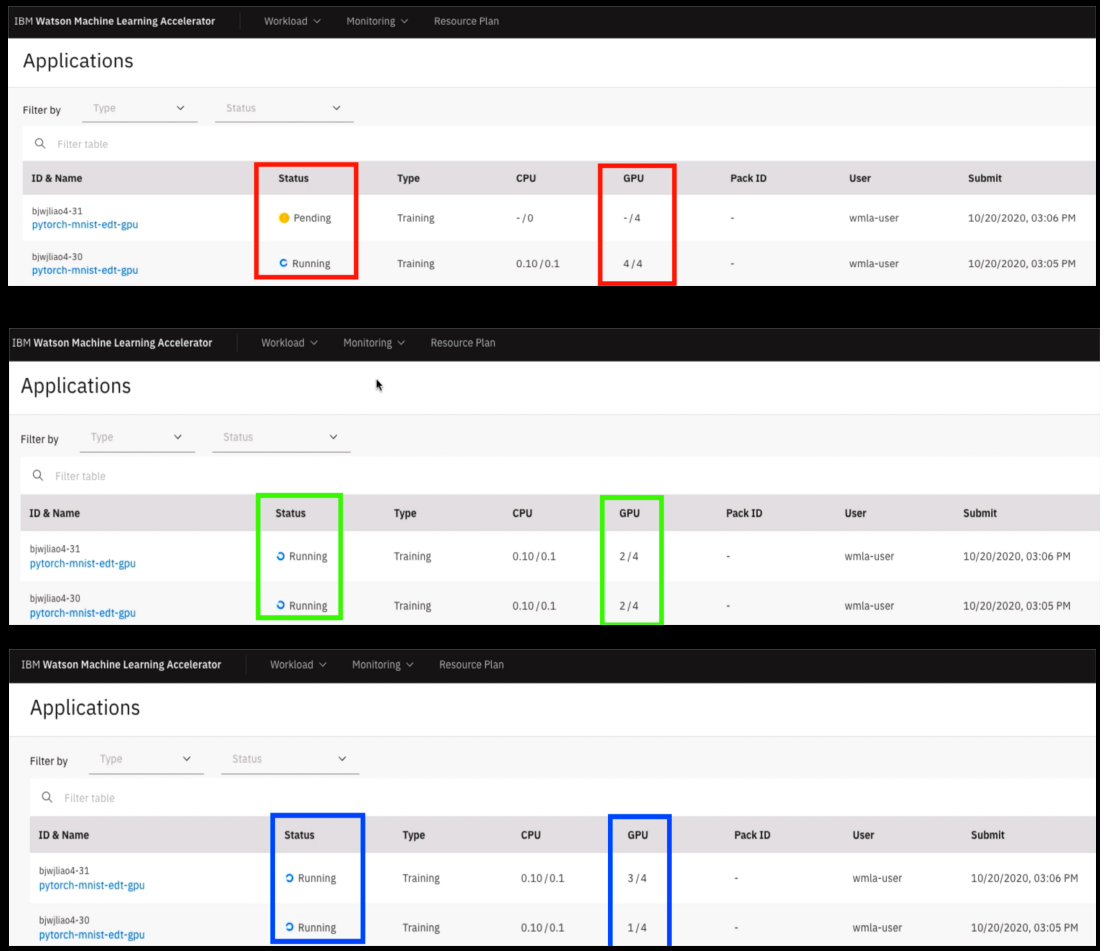
- Hadoop Execution Engine
  - HEE users can securely connect to Cloudera Data Platform 7.1.5 for data access and offloading workloads.
  - Support insert to code for HEE connectors
  - JupyterLab working with HEE connectors
  - Notebook running in spaces with HEE connectors



Watson Studio in Cloud Pak for Data 4.0

Elastic GPU Computing

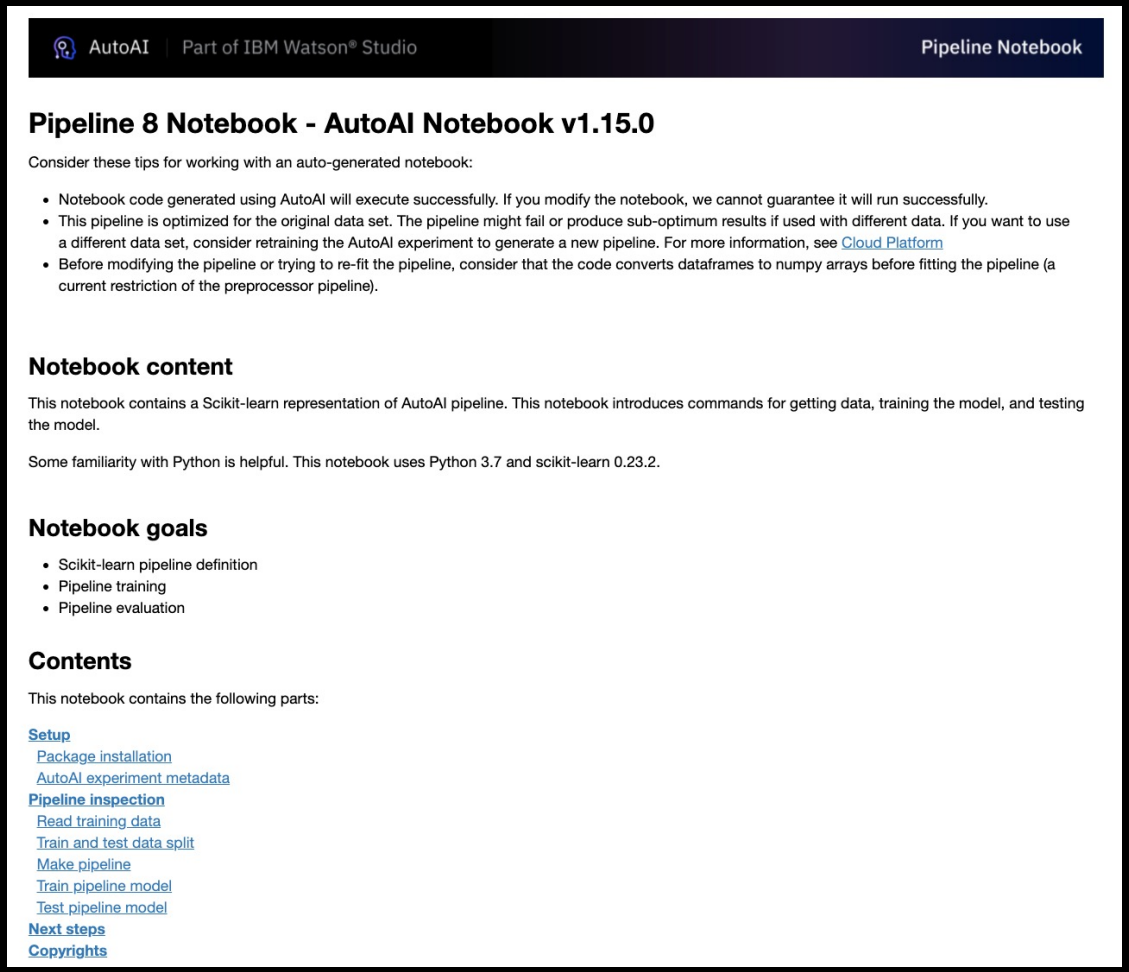
- Deep Learning libraries updates including:
  - NVIDIA CUDA Toolkit 11.0 that supports for NVIDIA A100 GPU
- High Availability with multiple active replicas
- Training Backup and Restore
- Auditing (Audit Logging)
- Openshift Container Storage 4.6.0
- Enhanced security with TLS 1.2.0



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



AutoAI | Part of IBM Watson® Studio — Pipeline Notebook

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.

Some familiarity with Python is helpful. This notebook uses Python 3.7 and scikit-learn 0.23.2.

Notebook goals

- Scikit-learn pipeline definition
- Pipeline training
- Pipeline evaluation

Contents

This notebook contains the following parts:

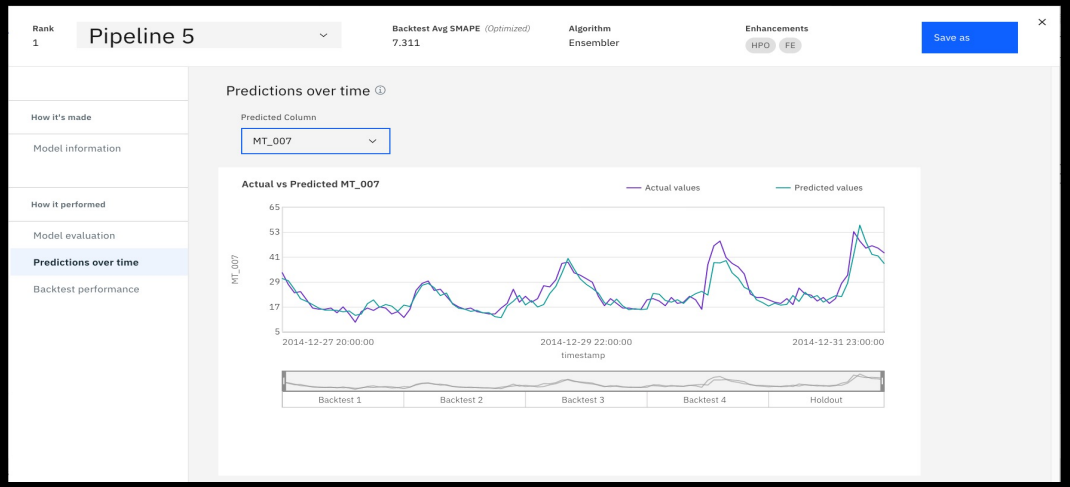
Setup
Package installation
AutoAI experiment metadata
Pipeline inspection
Read training data
Train and test data split
Make pipeline
Train pipeline model
Test pipeline model
Next steps
Copyrights

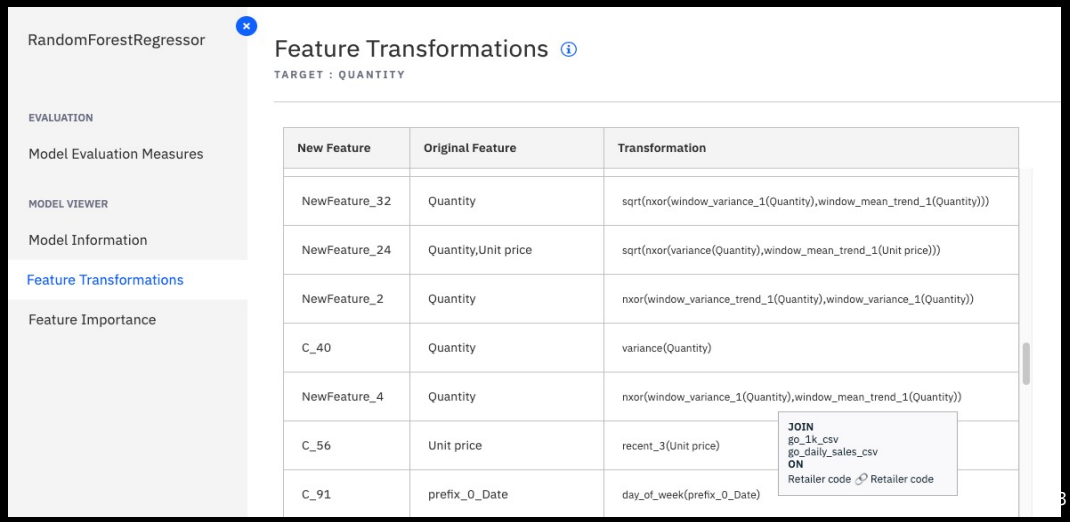
Beta Features In Watson Studio on Cloud Pak for Data 4.0

- **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.

Tests to be run on ML/DL models differ from straightforward application testing

- **Drift**: Any change in input data – also known as Drift – can cause the model to make inaccurate decisions, impacting business KPIs
- **Bias**: Training data may be cleaned to be free from bias, but runtime data may induce biased behavior of model
- **Explainability**: Traditional statistical models are simpler to interpret and explain
- **Missing Validation/Test Data**: Model training data sets may not capture the range of data or combinations that could be encountered in runtime

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

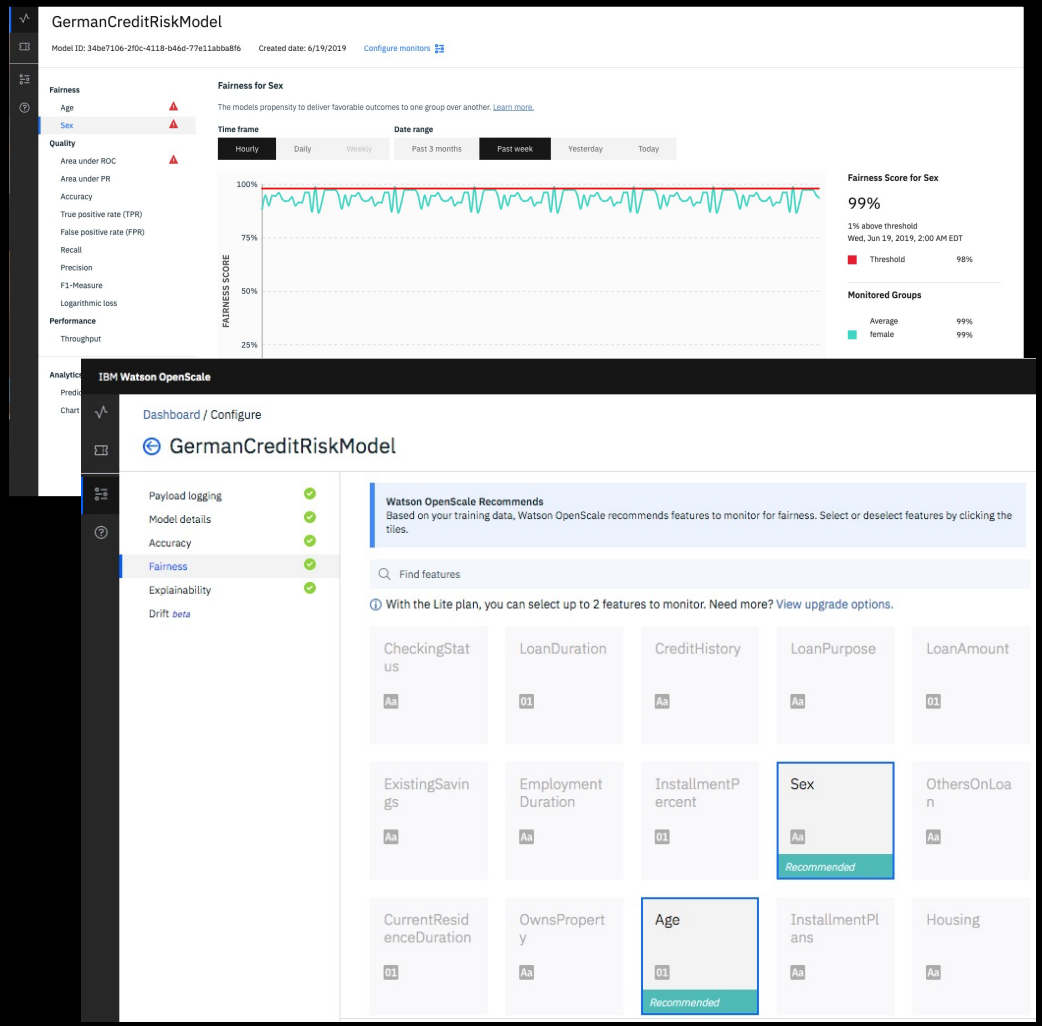
Align model performance with business outcomes

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

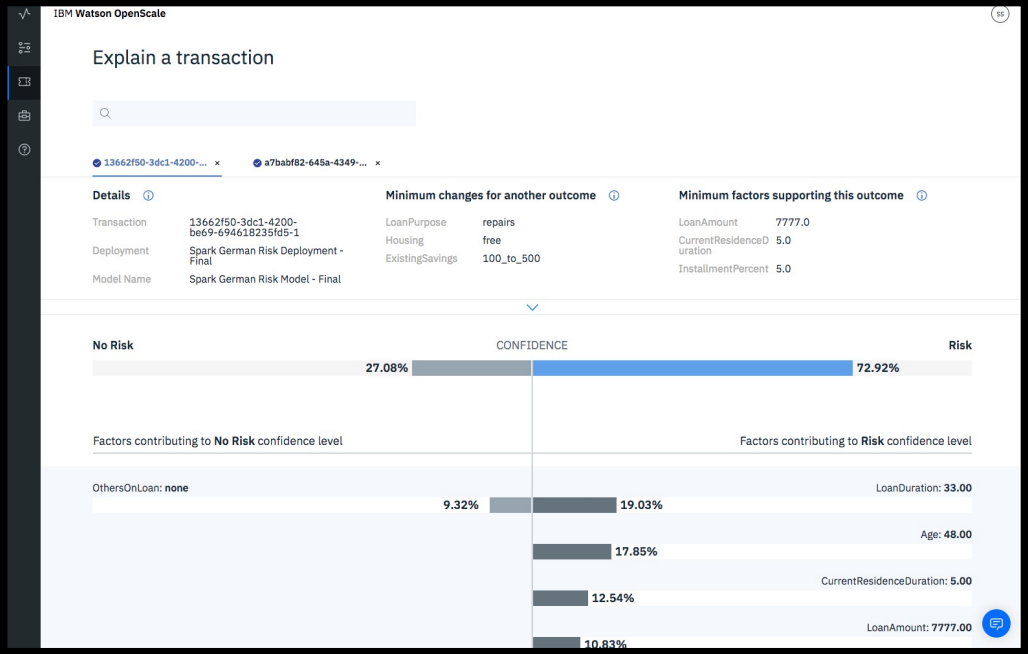
Explainability

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

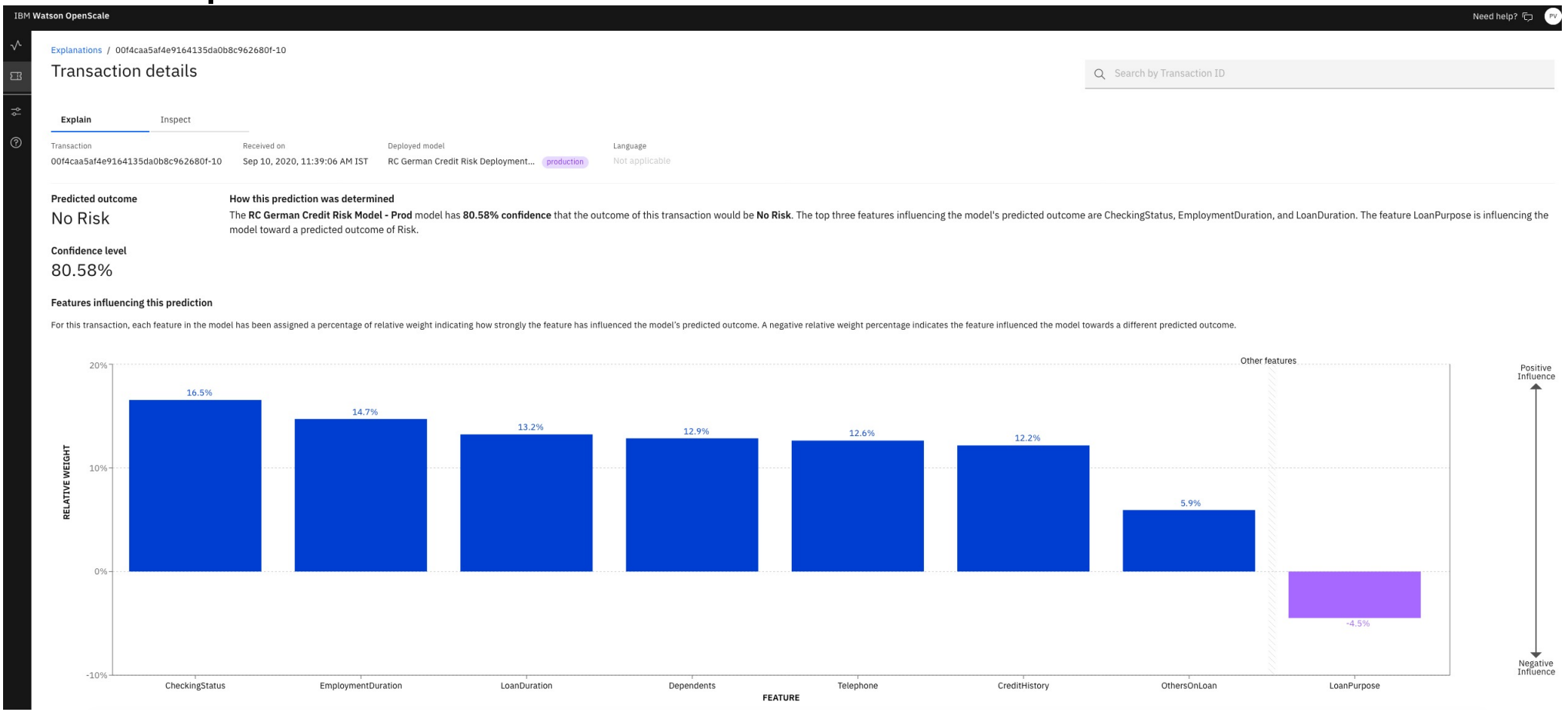
It provides a simple explanation that is user friendly and interactive

Value:

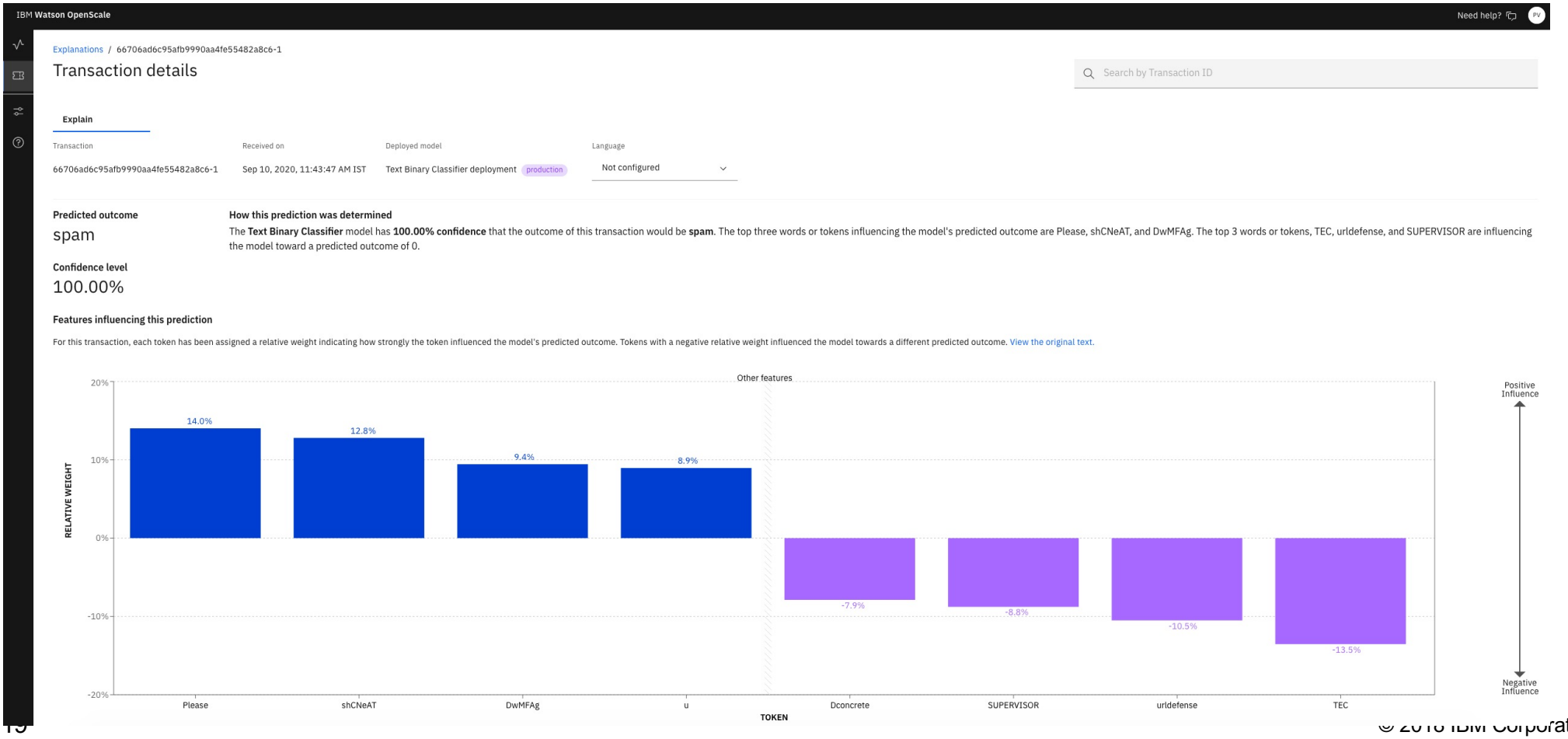
- Explain individual transaction level decisions made by the model in run time, including details about most important attributes and their values in order to assist in compliance and customer care situations
- Analyze individual transactions in a what-if manner in order to understand how model behavior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



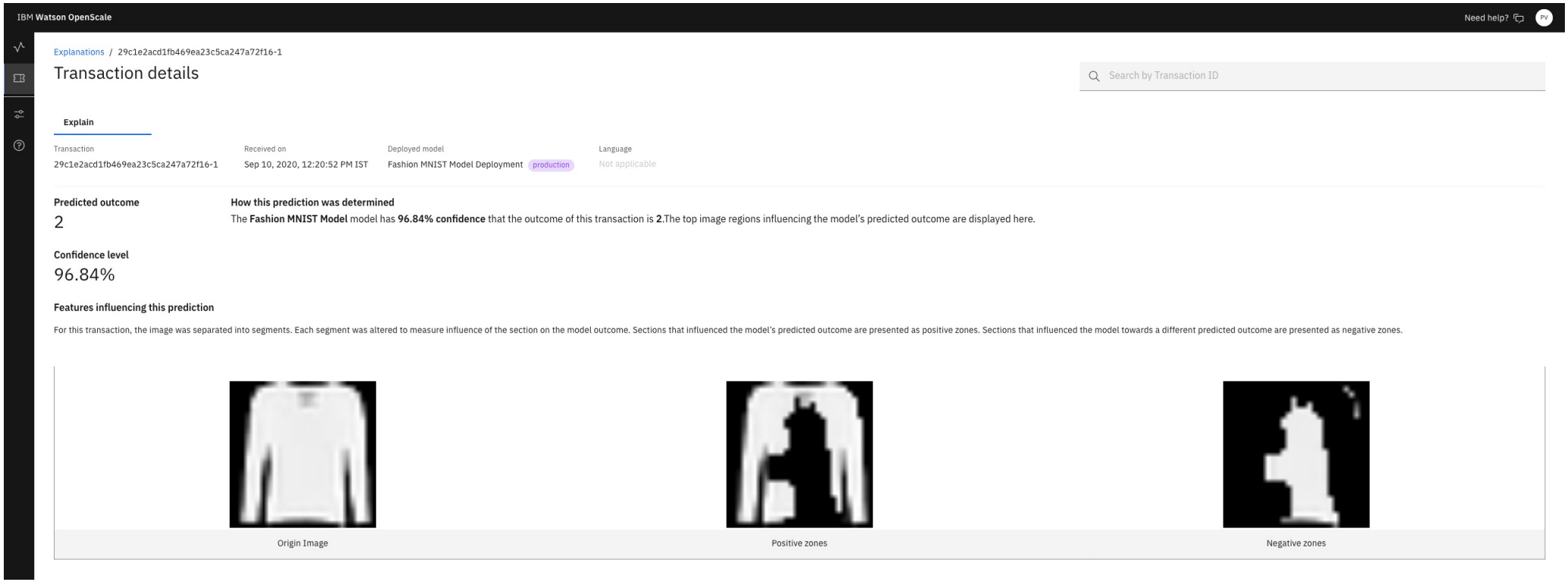
Lime Explanation for structured data



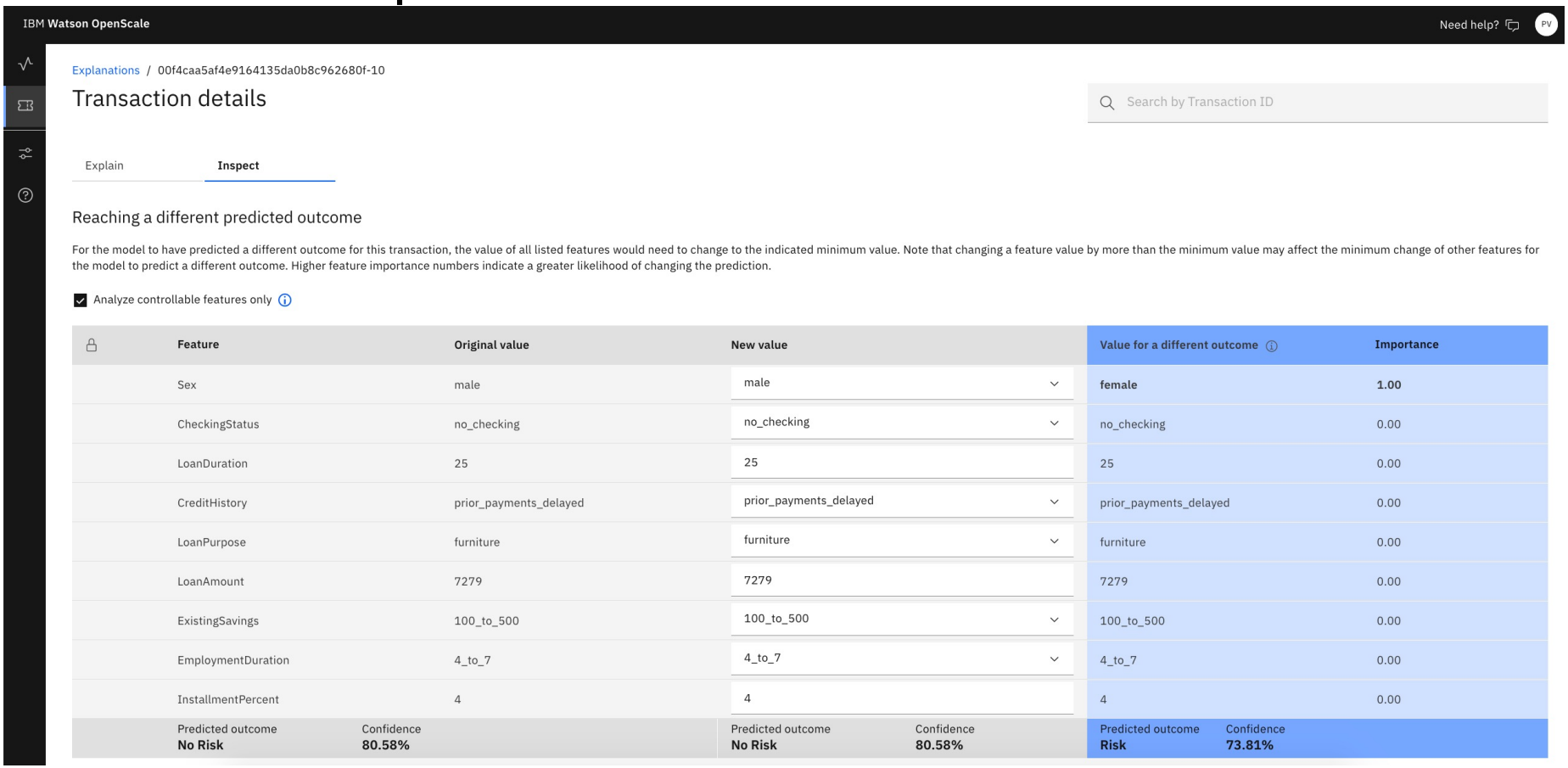
Explanation for unstructured text



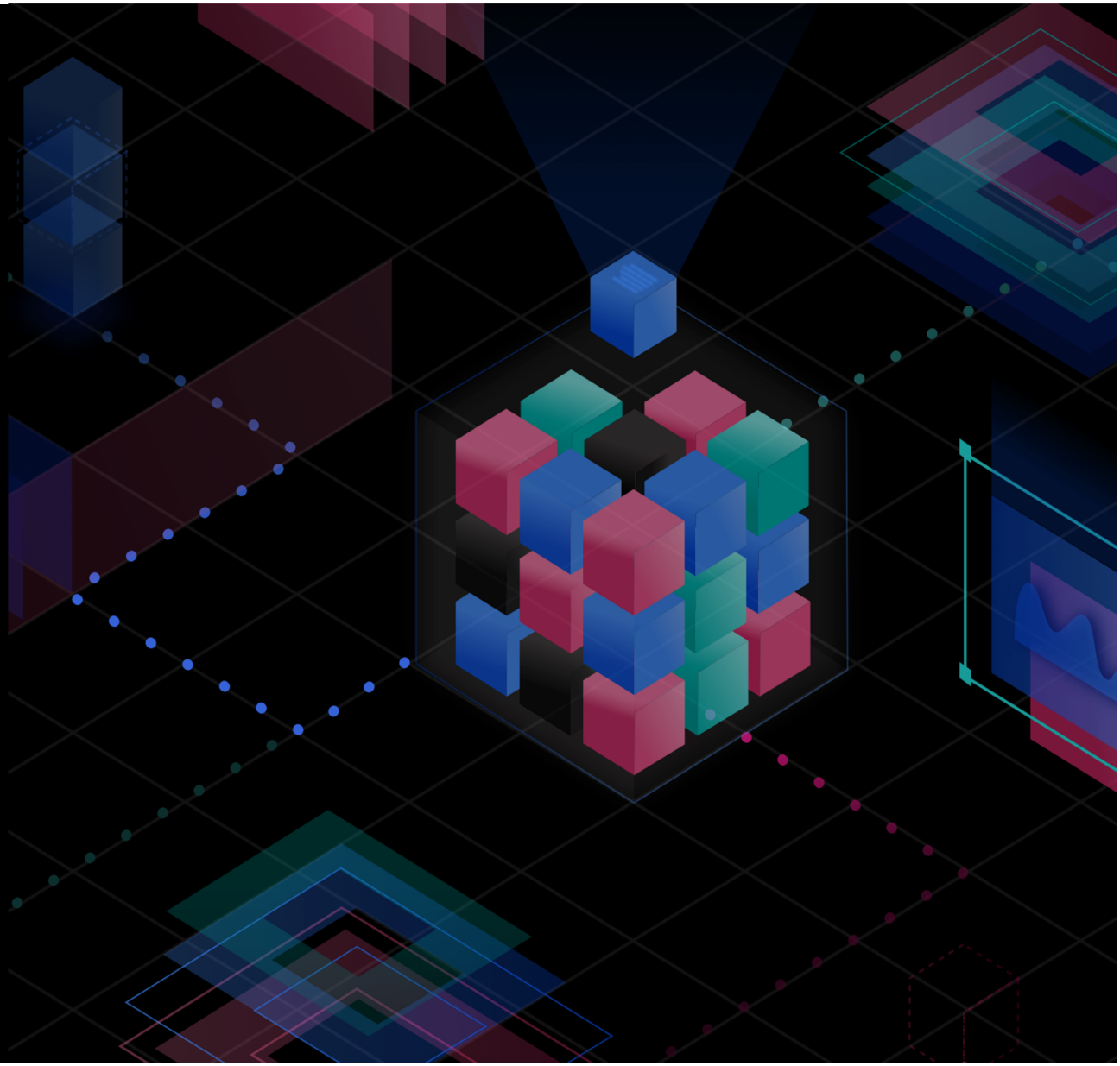
Explanation for unstructured image



Contrastive Explanation



Thanks



Rachana Vishwanathula
Hybrid Cloud Build Team
rachvis1@in.ibm.com