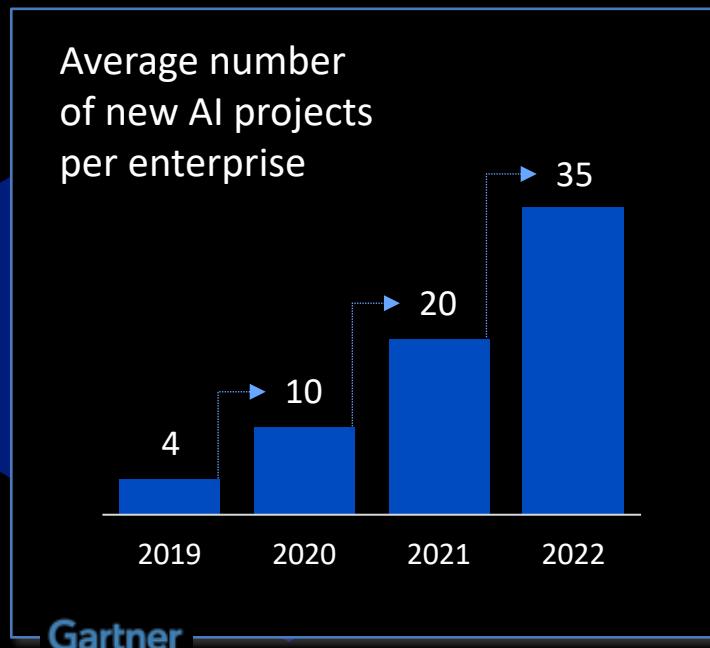
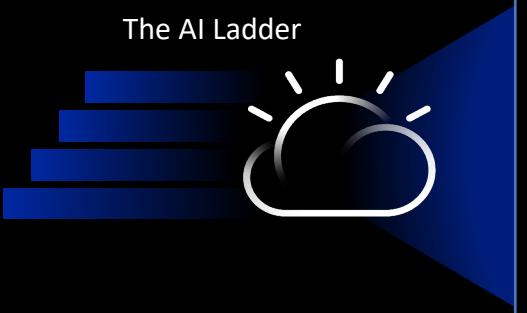


# Accelerate and Expand Use of Data Science and ML libraries with Anaconda Repository for IBM Cloud Pak for Data

**Greg Filla**, Senior Offering Manager, Watson Studio  
**Thomas Schaeck**, Distinguished Engineer, Watson Studio



# Enterprises are accelerating their journeys to AI



**21%** are making AI their largest tech investment

**57%** have built AI skilled teams of greater than 50

**2.5X** greater investment in AI lifecycle platforms

FORRESTER

# IBM delivers outcomes with clients

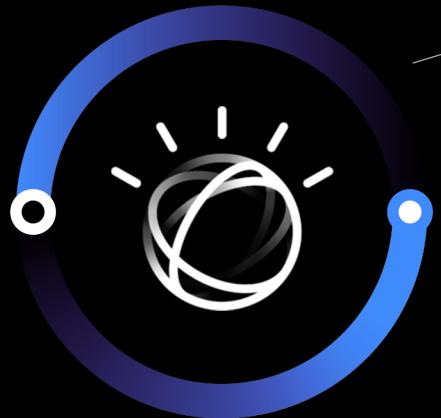


WUNDERMAN  
THOMPSON

Achieved **200%** improvement in AI  
customer prediction model

**ExxonMobil**

**40%** savings in time to prepare  
data to build AI models for 3D  
seismic maps



**ANYLINE**

**10x** jump in AI training productivity to build AI pipelines for  
smartphones to read, interpret visual information

**JPMorganChase**

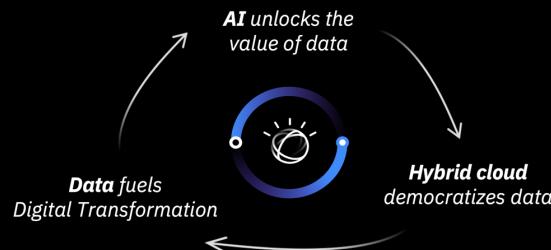
Deployed workflows to apply machine  
learning to AI-driven Quant Model Risk  
Management

**KPMG**

Enables trust in outcomes by automating  
model bias, regulatory change, etc. when  
approving a loan or line of credit.

# An information architecture for AI

Simplifies and automates the AI Ladder



One Platform,  
Any Cloud

## Watson AI Lifecycle and Applications

### Cloud Pak for Data

Pre-integrated Data, AI and Cloud services delivered within an open and extensible cloud native platform

Cloud Pak for Data components:

- Collect Data
- Organize Data
- Analyze & AI
- Infuse AI

Red Hat OpenShift

Supported clouds:

- IBM Cloud
- AWS
- Azure
- Google Cloud
- Hyperconverged Private Cloud Systems and AI Accelerated Infrastructures

# IBM Data Science and Machine Learning Portfolio

Prepare and Organize Data

Build and Train AI Models

Deploy and Run AI Models

Manage and Operate Trusted AI



AutoAI Lifecycle Automation and Governance



Cloud Pak for Data



IBM Cloud



Hyperconverged Private  
Cloud System

# IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business



## Analyze any data, no matter where it lives

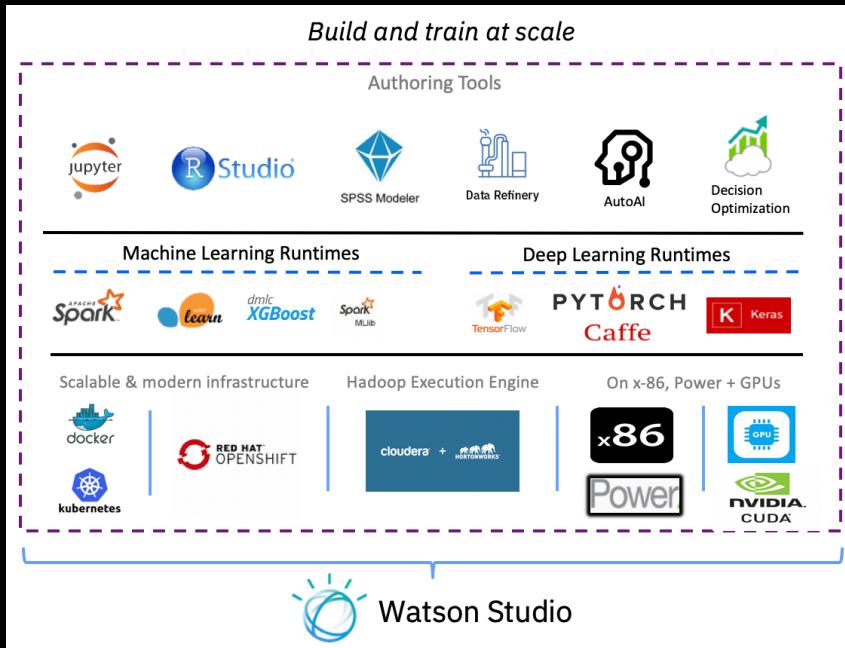
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

## Empower your entire organization with notebooks, visual productivity, and automation tools

Leverage your entire organization with a variety of tools in a single integrated platform

## One platform to rule them all from discovery to production

Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



# IBM partners with Anaconda to bring comprehensive, curated open source libraries to a multicloud, governed data and AI platform



IBM and Anaconda teaming to ease and expand **open source data science deployment into an enterprise**. To address the growing needs of deploying data science projects **across any cloud**, IBM and Anaconda bring together the power of open source innovation and enterprise AI deployment **at scale**. With more than 20 million users Anaconda is the world's most popular data science tool and the foundation of modern machine learning. This unique partnership is aimed at accelerating time to build and scale AI while taking advantage of **vibrant, open source innovation**.

# Benefits of Anaconda Repository for Cloud Pak for Data

*Harness open-source building blocks for real data science.*

## Summary

- Cutting-edge data science and machine learning technology comes from open-source packages and libraries
- Unmanaged open-source technology comes with risk
- *Anaconda Respotiory* solves for that, enabling you to curate, block, and blacklist packages based on CVE scores, alerts, and reports.

## Features

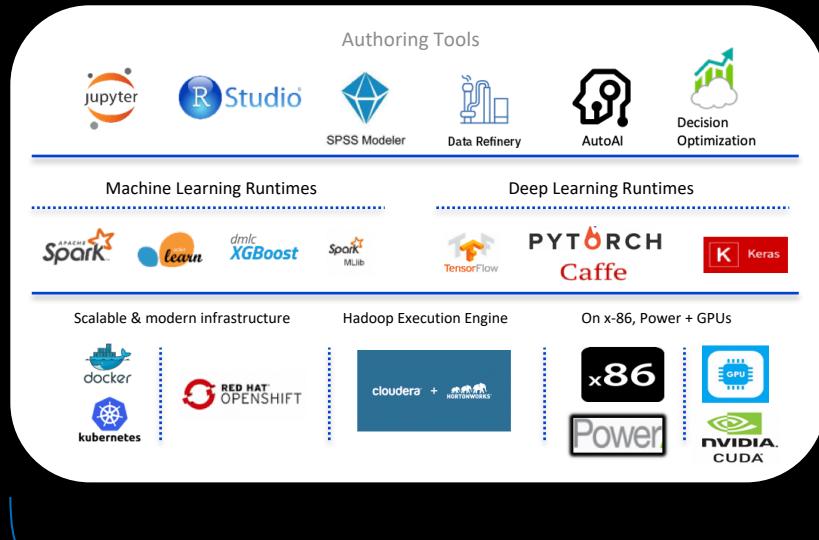
- Clearinghouse for build artifacts (packages and libraries) - along with their metadata
- Comprehensive history of repository events to ensure governance and security
- Supports distribution of consumable artifacts to end users, package managers, and CI servers
- Scan open source packages for security

<b>Innovation</b> 	<b>Management</b> 	<b>Security</b> 
Stay current with the latest from the opensource brain trust	Control the who, what, where, when and how	Catch vulnerabilities before they catch you

# Better Together

## Anaconda Repository + Watson Studio in Cloud Pak for Data

Build and train at scale



Watson Studio

- Data Scientists use Jupyter notebooks/Lab JupyterLab with managed environments as they do today
- With Anaconda Repository, Cloud Pak for Data users can govern these open source packages more readily
- New packages required by data scientists will go through admin approval
- Premium Services/Support – scan open source packages for security



# Watson Studio Notebook User Experience With Anaconda Repository

## Cloud Pak for Data administrators:

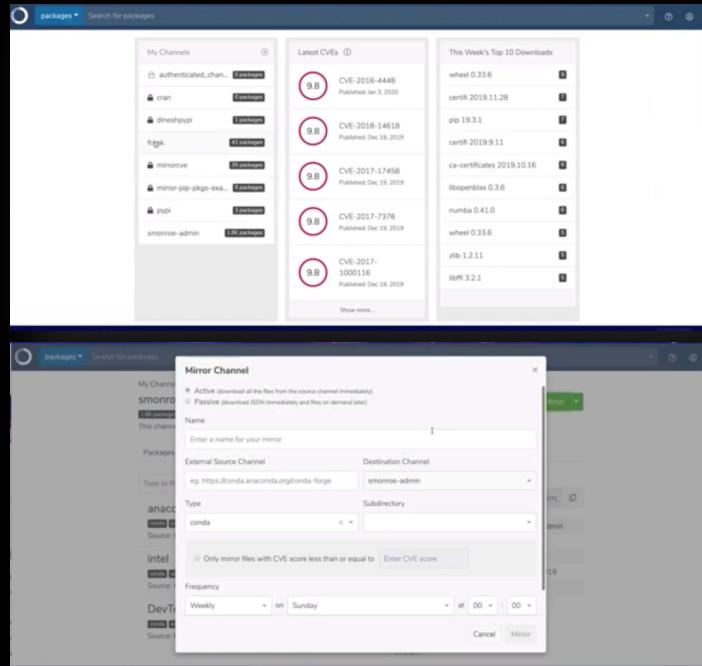
- Configures Conda channels and artifacts available for inclusion in Watson Studio environments
- Can mirror channels/repositories from cloud for installing in air-gapped environment

## Data scientist users:

- Adds a notebook asset to a project
- Selects the environment runtime to execute the notebook (defines hardware and software configuration)
- Selected environment contains the approved packages

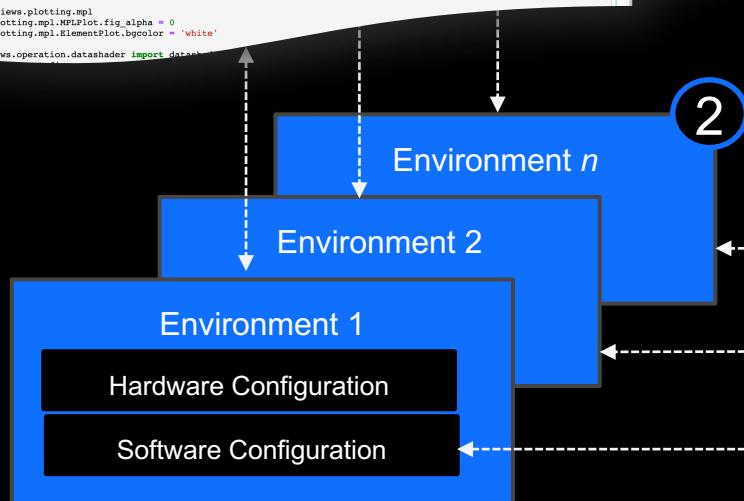
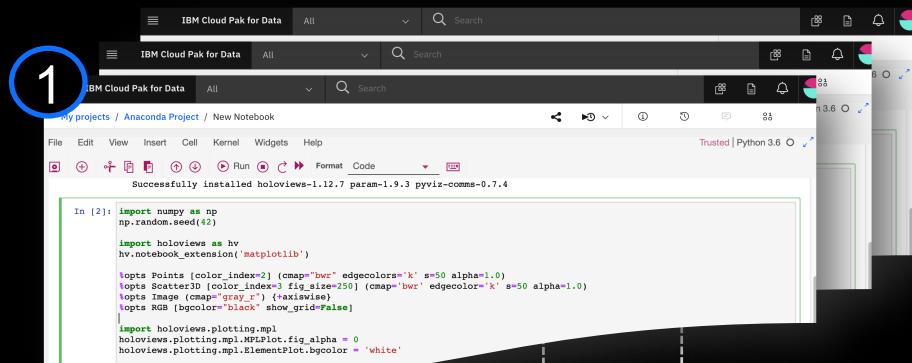
## User benefits:

- Administrators have **granular control** over which packages are available for data scientists
- Data Scientists have all dependencies they require
- Administrators can **securely make new packages available**

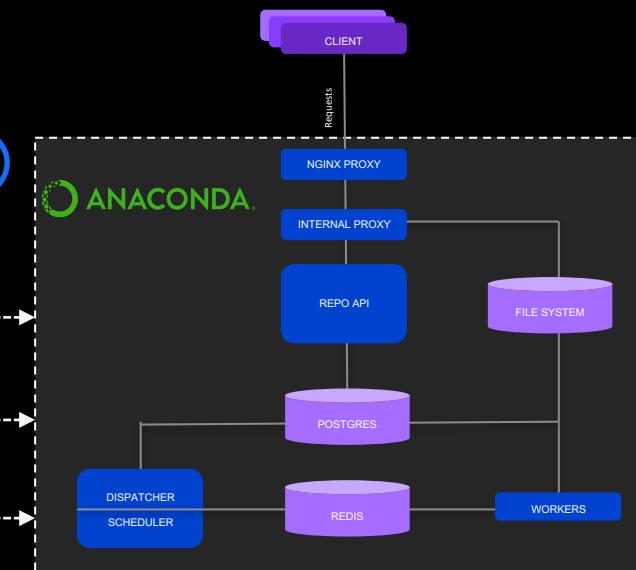


Anaconda Team Repository Admin View

# How Anaconda Repository for Cloud Pak for Data works



1. Watson Studio Notebook
2. Watson Studio Environment
3. Anaconda Repository



# Using Anaconda Repository with IBM Cloud Pak for Data



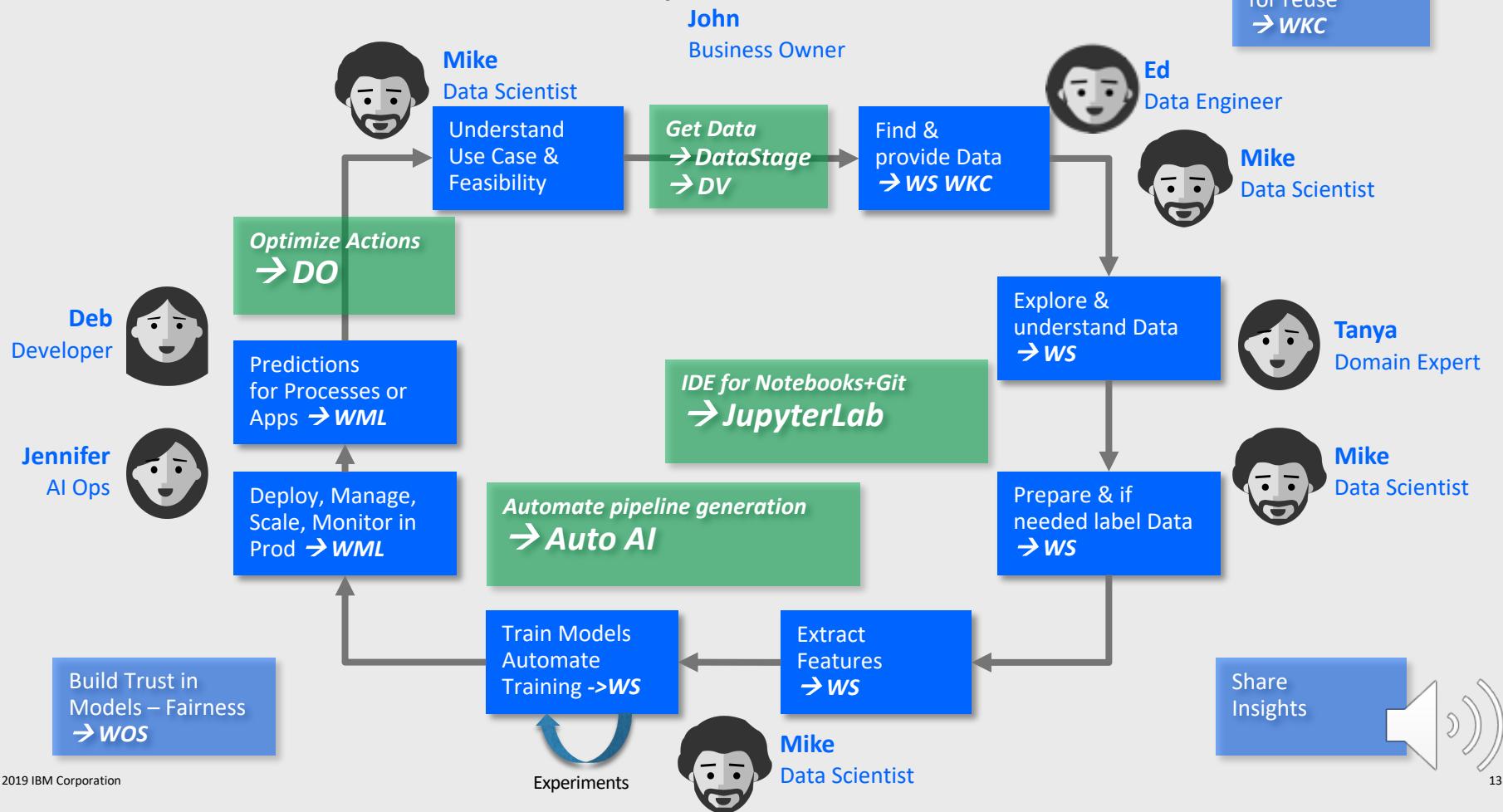
Thomas Schaeck

Distinguished Engineer, Watson Studio and Integration

IBM Data and AI

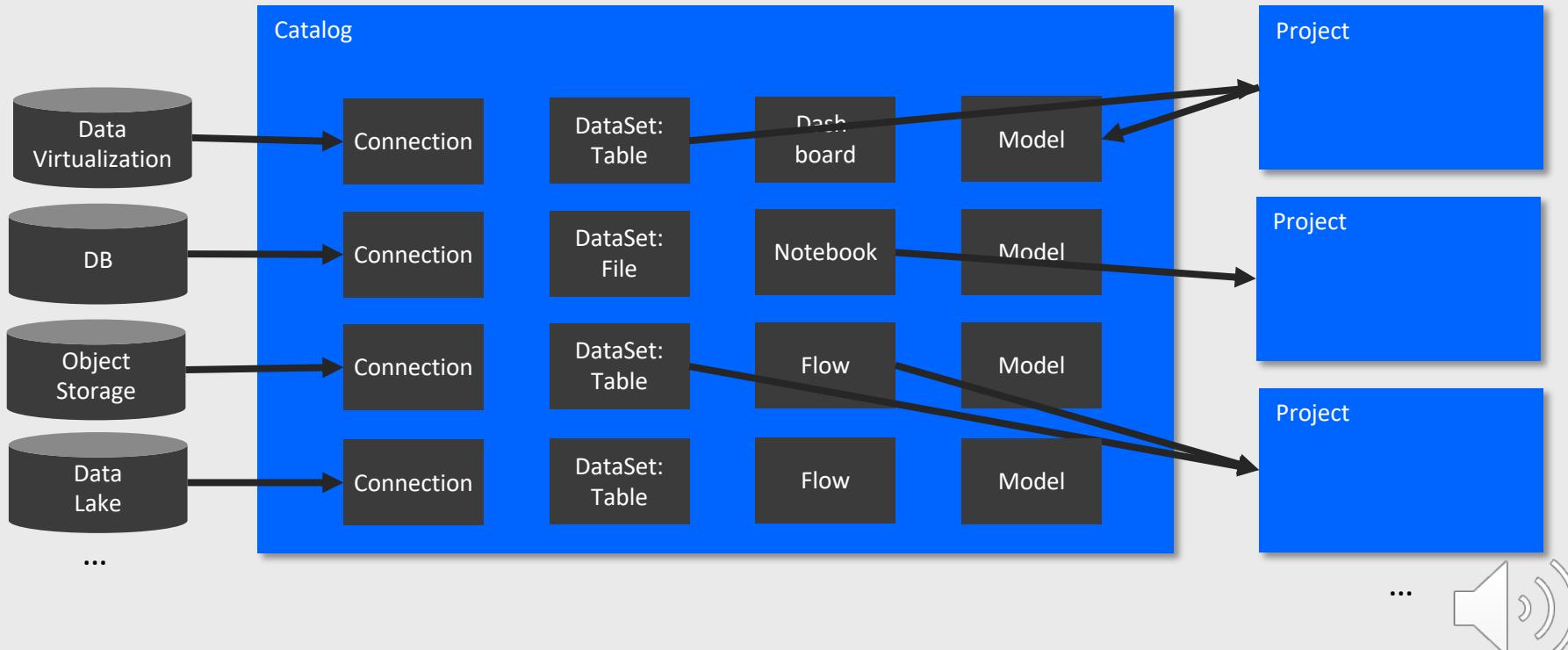


# End to End Data Science / ML Cycle

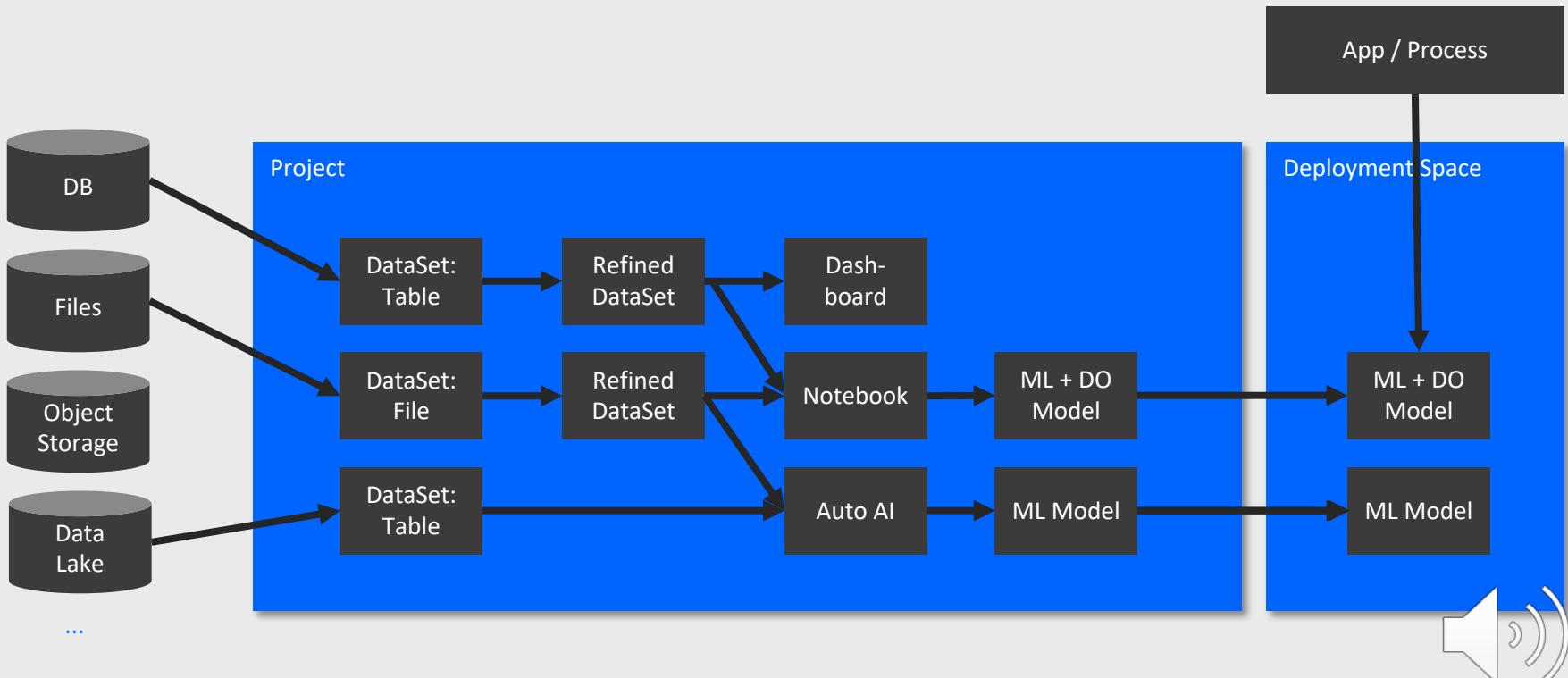


# Catalog Data from many Sources → Use in Projects

Supporting wide range of data sources, the example below just illustrates some



# Use Models from Apps and/or Business Processes



# Powering Projects: Environments Runtimes

Project

Environment

Environment

Environment

Environment

Environment

- **Common Runtime Environments Service** used for JupyterLab, Jupyter, RStudio, Refinery
- **Per Project - Per User - Per Environment Type Containers**
- **Reservations for #cores #GB per default**
- **Environment Customization with Conda**
- **Automatic Clean Up** of idle environments after N hours of inactivity if data scientists don't terminate their environments



# Existing use of Conda and Anaconda Open Source in Cloud Pak for Data

- Default Environment images for Jupyter / JupyterLab use Anaconda OS
- Environment customization allows to add extra Libs from Conda Channels

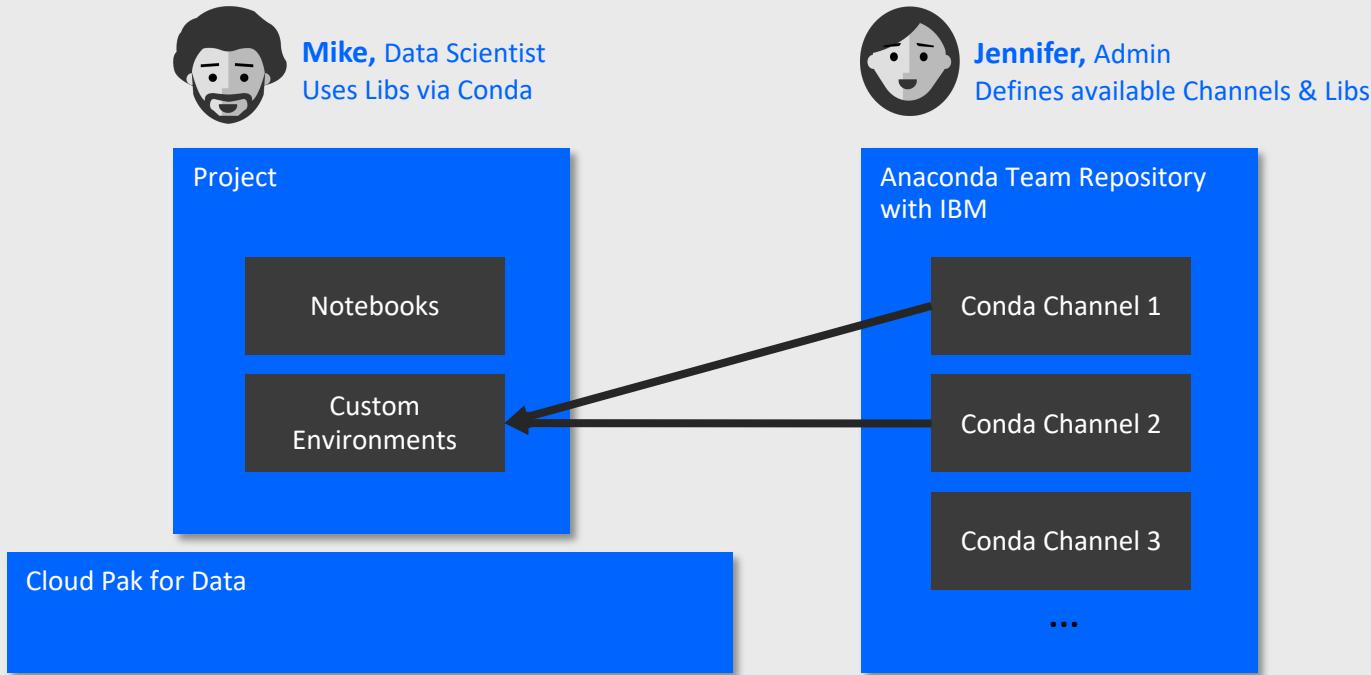


# Need for Enterprise Control, Speed and Independence

- Many enterprises need control of what libs their data scientists use
  - Loading libs and dependencies from external Internet channels can be slow
  - Dependency on external Internet channels not always acceptable
- IBM and Anaconda partnered to bring Anaconda Team Repository with IBM
- Functionally identical with Anaconda Team Edition
  - Provides fast library loading from local repo proxy service
  - Channels are enterprise managed: Control what libs can be used
  - Configure Cloud Pak for Data to use desired Conda Channels



# Customize Environments using Libs from Conda Channels



Enterprises may block connectivity from Runtime Environments on Cloud Pak for Data to external libs in order to force loading all libs from Conda Channels managed using Anaconda Team Repository with IBM



# Configuring Cloud Pak for Data to use Conda channels from Anaconda Repository with IBM

Customize `/user-home/_global_/config/conda/.condarc` like this

```
channel_alias: http://<server>/api/repo
channels:
- forge
default_channels:
- http://<server>/api/repo/forge
```





packages ▾

Search for packages



## forge / gama ←

conda 12 files Latest: 2.06

This package has no description.

Files 12

Dependencies 1

Dependants 0

Type to filter

Sort

Version ↴

Name	Version	Platform	Score	CVEs	Size	Uploaded
gama-2.06-h0a44026_0.tar.bz2	2.06	apple osx-64	--	--	1013.94 KB	4/2/20 4:27 PM
gama-2.06-he025d50_0.tar.bz2	2.06	Microsoft win-64	--	--	602.11 KB	4/2/20 3:17 PM
gama-2.06-hf484d3e_0.tar.bz2	2.06	Delta linux-64	--	--	1.07 MB	4/2/20 2:11 PM
gama-2.03-h0a44026_0.tar.bz2	2.03	apple osx-64	--	--	969.93 KB	4/2/20 4:27 PM
gama-2.03-he025d50_0.tar.bz2	2.03	Microsoft win-64	--	--	557.45 KB	4/2/20 3:17 PM
gama-2.03-hf484d3e_0.tar.bz2	2.03	Delta linux-64	--	--	1.12 MB	4/2/20 2:11 PM

1 - 12 of 12

## Install

conda install -c forge gama



Latest 2.06

Version:

Last Apr 2, 2020

Published:

License: GPL-3.0

Family: conda

Platforms: apple osx-64



linux-64





# New environment

## Define environment details

Name

Thomas Test Environment

Description (optional)

Test

## Define configuration

Type

 Default  Spark  GPU  Hadoop

Hardware configuration

Specify: 2 vCPU and 4 GB RAM



Software version

Default Python 3.6



Cancel

Create



# Thomas Test Environment

Test

Last updated 5/25/2020, 4:36 PM by admin

## Summary

New notebook +

Environment	Thomas Test Environment
Creator	admin
Tool	Notebook
Language	Python 3.6
Hardware configuration	Specify: 2 vCPU and 4 GB RAM
Software configuration	Default Python 3.6

## Software configuration details

- \_libgcc\_mutex=0.1
- \_py-xgboost-mutex=2.0
- pytorch\_collect=1.0

## Customization

You do not have a customization yet. [Create one now.](#)





## Software configuration details

```
-_libgcc_mutex=0.1
-_py-xgboost-mutex=2.0
-_pytorch_select=1.0
-absl-py=0.7.1
-alabaster=0.7.12
-arrow-cpp=0.11.1
-asn1crypto=0.24.0
-astor=0.7.1
-astroid=2.1.0
-astropy=3.1.1
-atomicwrites=1.3.0
-attrs=18.2.0
-avrouiwidget=0.12.9
-babel=2.6.0
-backcall=0.1.0
-backports=1.0
-backports.os=0.1.1
-backports.shutil_get_terminal_size=1.0.0
-beautifulsoup4=4.7.1
-biopython=1.72
-bitarray=0.8.3
-bkcharts=0.2
-blas=1.0
-blaze=0.11.3
-bleach=3.1.0
-blosc=1.15.0
-bokeh=1.0.4
-boost=1.67.0
-boto=2.49.0
-boto3=1.9.82
-botocore=1.12.82
-bottleneck=1.2.1
-bzip2=1.0.8
-c-ares=1.15.0
-ca-certificates=2020.1.1
-caffe-base=1.0_1.6.2
-cairo=1.14.12
-certifi=
-cffi=1.12.3
-chardet=3.0.4
-click=7.0
-cloudpickle=0.7.0
-clyent=1.2.2
-colorama=0.4.1
-colour=0.1.5
-contextlib2=0.5.5
```

## Customization

```
# Modify the following content to add a software customization to an
environment.
# To remove an existing customization, delete the entire content and click
Apply.

# Add conda channels below defaults, indented by two spaces and a hyphen.
channels:
    - nodefaults ←
    - forge ←

# To add packages through conda or pip, remove the comment on the following
line.
# dependencies:

# Add conda packages here, indented by two spaces and a hyphen.
# Remove the comment on the following line and replace sample package name with
your package name:
#   - a_conda_package=1.0
#   - gama=2.0.6 ←

# Add pip packages here, indented by four spaces and a hyphen.
# Remove the comments on the following lines and replace sample package name
```



IBM Cloud Pak for Data

All

Search

My projects / ATE / Test Notebook

File Edit View Insert Cell Kernel Widgets Help

Not Connected Trusted Python 3.6

In [\*]: `from gama import GamaClassifier`

In [ ]:

Information

General Environment

Environment definition

Thomas Test Environment

Language Python 3.6

Hardware configuration 2 vCPU, 4 GB RAM

Software configuration [View details](#)

Runtime status Running

Open terminal

Speaker icon

This screenshot shows the IBM Cloud Pak for Data Jupyter Notebook interface. The top navigation bar includes 'IBM Cloud Pak for Data', 'All', 'Search', and various system icons. The main menu bar has options like File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. The toolbar below the menu includes icons for file operations, run cells, and format/code selection. In the code editor, the first cell contains the Python code 'from gama import GamaClassifier'. The second cell is currently empty, indicated by 'In [ ]:'. To the right, the 'Information' panel is open, showing the 'Environment' tab selected. It displays the environment definition as 'Thomas Test Environment', the language as 'Python 3.6', hardware configuration as '2 vCPU, 4 GB RAM', and software configuration with a link to 'View details'. The runtime status is shown as 'Running'. At the bottom right, there is a speaker icon.

## Next steps:

Tap into IBM resources and expertise  
to help drive your business outcomes

Test drive IBM Cloud Pak  
for Data for free with a  
7-day trial. No  
installation required

Start your trial now

[http://ibm.biz/cloudpakdatatrial](https://ibm.biz/cloudpakdatatrial)

See how to sync  
ModelOps to DevOps to  
accelerate AI ROI  
  
Read the analyst report

<https://ibm.biz/Bdqzkp>

Engage with our experts  
  
Connect with our Data  
Science & AI Elite team  
for a customized 4-6  
weeks AI consultation

<https://ibm.biz/Bdqzkn>

Try the IBM AI lifecycle  
portfolio  
  
Start your free trial of  
IBM Watson Studio,  
Watson Machine  
Learning and Watson  
Knowledge Catalog

<https://ibm.biz/Bdqzkf>

# Thank You

**Greg Filla**, Senior Offering Manager, Watson Studio

**Thomas Schaeck**, Distinguished Engineer, Watson Studio

