

Digital Developer Conference

# Data & AI

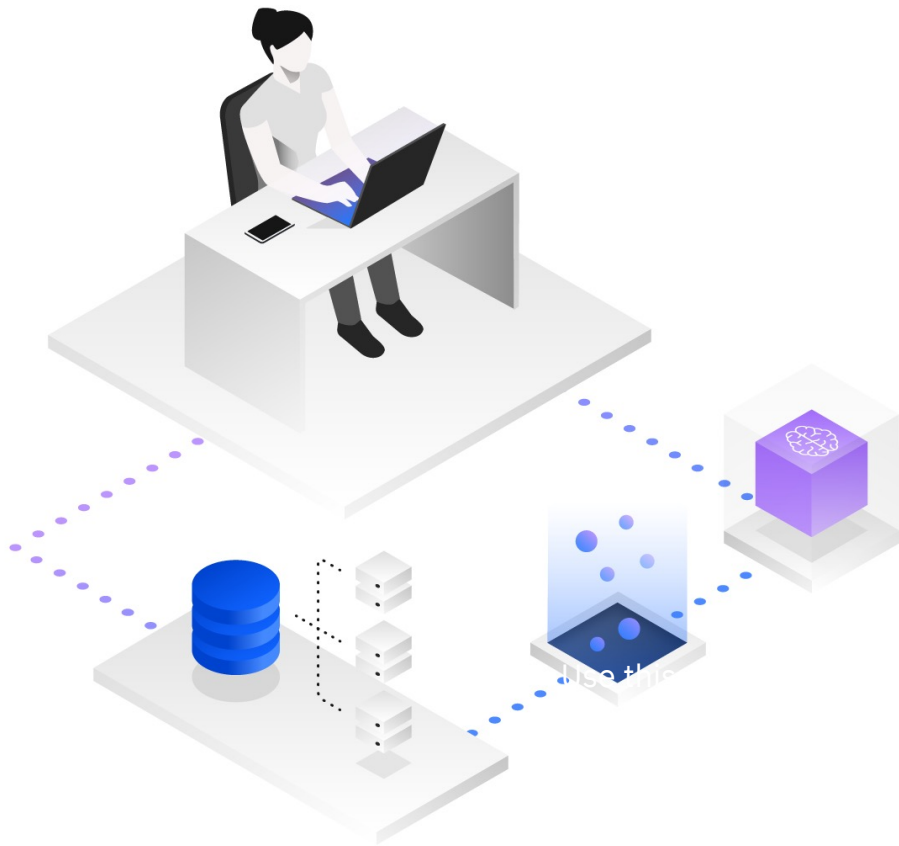
## Lab: Automating your data science Jupyter notebook workflow with open source tools

**Speakers:** Yiwen Li  
Patrick Titzler

**Agenda:**

1. AI extensions for JupyterLab
2. Open data sets

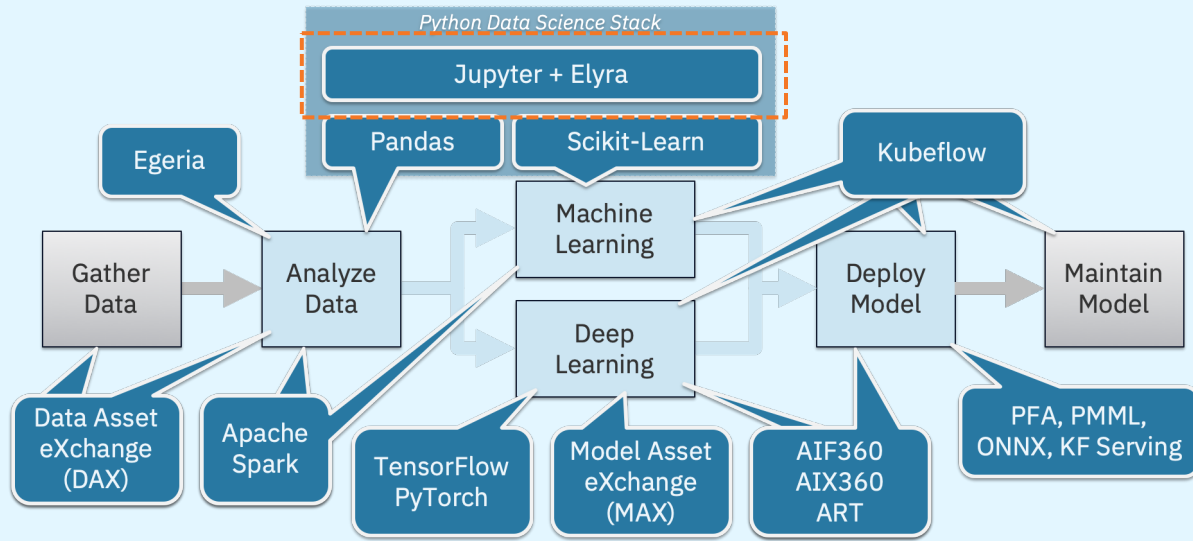
<https://github.com/CODAIT/DDC-data-and-ai-2021-automate-using-open-source>





- CODAIT aims to make AI solutions dramatically easier to create, deploy, and manage in the enterprise.
- 40+ developers/data scientists
- We contribute to and advocate for the open-source technologies that are foundational to IBM's AI offerings.

## Improving the Enterprise AI Lifecycle in Open Source



# Elyra: AI-centric extensions to JupyterLab

re-use code

Code snippets

source control

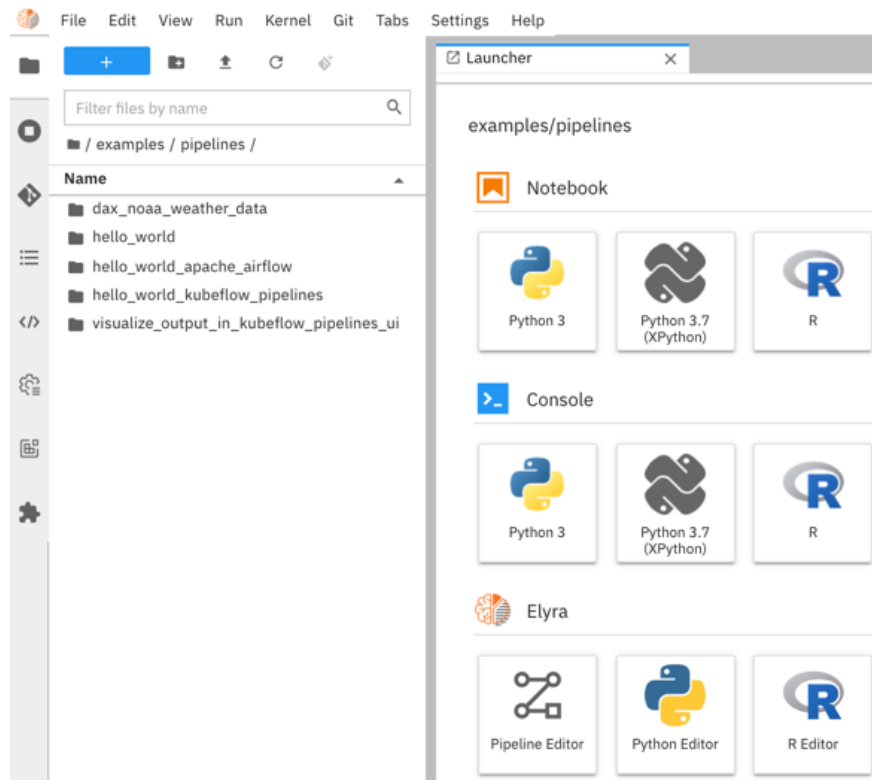
Git integration

Debug and run remotely

Python/R/notebook editor

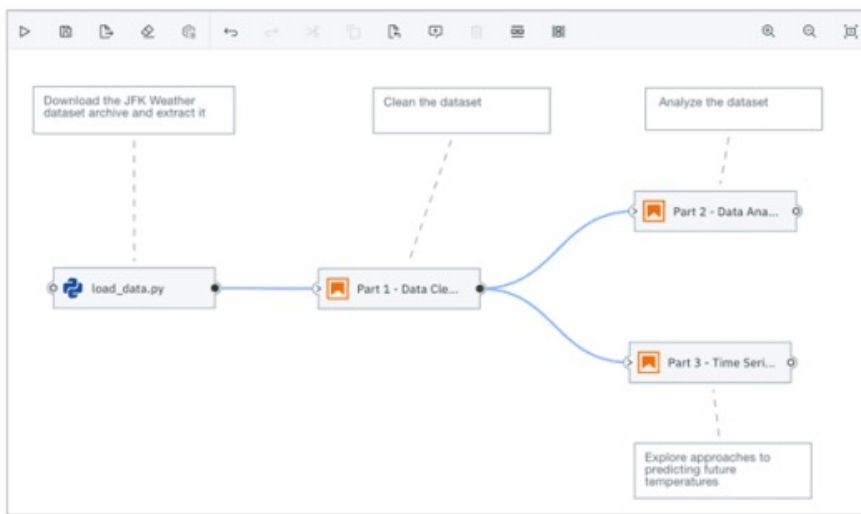
Create and run ML workflows

Visual pipeline editor



# Implementing ML workflows using pipelines

- Modular notebooks (or Python/R scripts) allow for re-use in other projects
  - Example: load data, cleanse data, train model, ...
- Assemble pipeline using Visual Pipeline Editor
- Run locally in JupyterLab, or remotely on Kubeflow Pipelines or Apache Airflow



Local execution  
in JupyterLab

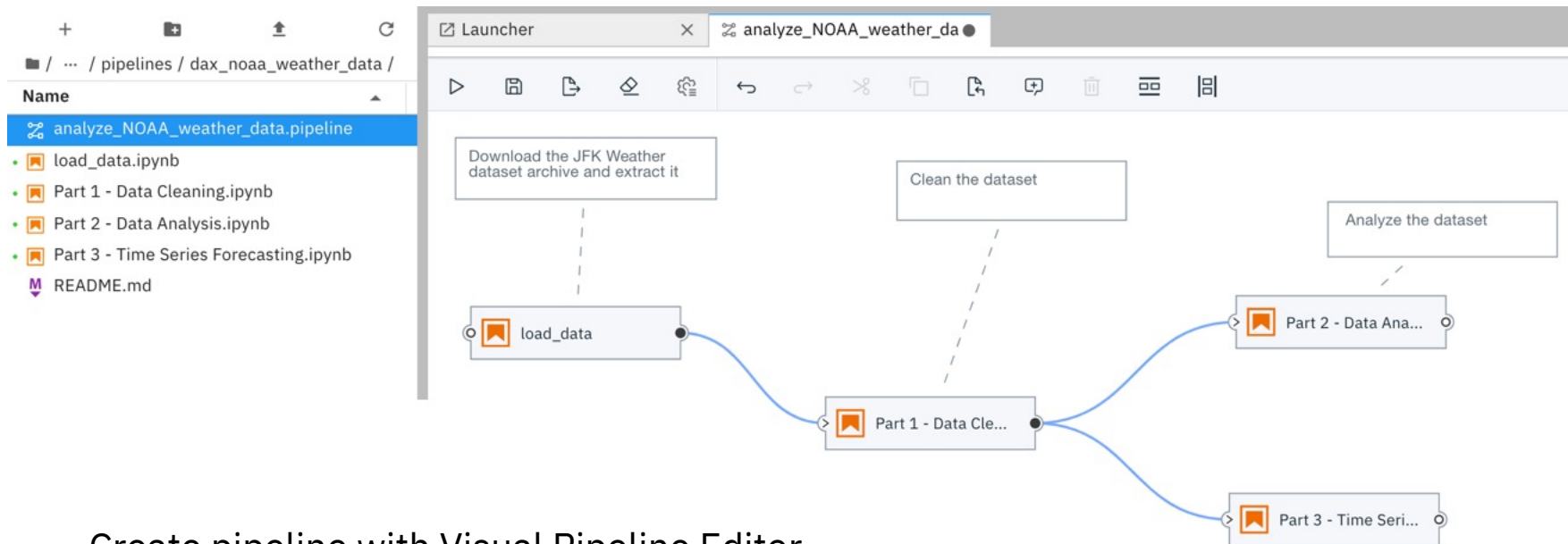


Kubeflow Pipelines



Apache Airflow

# Demo: Implementing an ML workflow using Elyra



- Create pipeline with Visual Pipeline Editor
- Run pipeline locally in JupyterLab
- Run pipeline on Kubeflow Pipelines
- [Tutorials](#)

# Data Asset eXchange

Data Asset Exchange offers high-quality datasets with clearly-defined open data licenses in standardized formats, according to IBM.

- Vetted data.
- Exclusive access to IBM Research datasets that have been used in creating popular AI products like [Debater System](#), Entity Recognition, and so on.
- Datasets with open data licenses for both business applications and advancing core science.
- Packaged with tutorials that shows how to read and analyze data. As well as, train machine or deep learning models on IBM Cloud using IBM Cloud AI services as well as multi-cloud AI open-sourced tools.

[ibm.biz/data-exchange](https://ibm.biz/data-exchange)

C&CS/ May, 2021 / © 2021 IBM Corporation

## Data Asset eXchange

Explore useful and relevant data sets for enterprise data science

[Learn More](#)



[What's New](#)



[Get Involved](#)



Dataset | CSV

NOAA Weather Data -  
JFK Airport

September 12, 2019



Dataset | IOB format

Groningen Meaning  
Bank - Modified

May 14, 2020



Dataset | CSV

Fashion-MNIST

September 12, 2019



Dataset | JPG, JSON

PubLayNet

October 25, 2019



Dataset | WAV

TensorFlow Speech  
Commands

March 17, 2020



Dataset | PNG, JSON

PubTabNet

November 11, 2019



# Data Preview and Data Glossary

DAX Dataset Preview

Dataset Metadata

Dataset Preview

Dataset Glossary

Image

Annotated Image (Generated with Notebook)

JSON

Image

Annotations

Categories

Annotations

Categories

- Explore the data

DAX Dataset Preview

Dataset Metadata

Dataset Preview

Dataset Glossary

## PubLayNet

- Understand the data


Feature	Description
images	JSON field containing a list of images and their metadata (size, ID, name)
annotations	Each object instance annotation contains a series of fields, including the category id and segmentation mask of the object.
annotations -> segmentations	Contains the polygon coordinates for the segmentation mask for the specific class instance (table, list, text etc)
annotations -> bbox	Contains the bounding box coordinates for the specific class instance (table, list, text etc).
annotations -> is_crowd	This field indicates whether the class instance is a single object (is_crowd=0) or multiple objects (is_crowd=1). In this dataset we only have single objects so this field is always set to 0.
annotations -> category_id	The class label for the current class instance. This indicates what the current bbox/segmentation mask encapsulates (table, list, text etc).
categories	JSON field containing a list of classes and their metadata (ID, name) This dataset has 5 categories (w/ corresponding "ids") - text ("1"), list ("2"), table ("3"), figure ("4"), figure ("5").

# Access data set notebooks in Watson Studio

IBM Cloud Pak for Data

Log In

Sign Up

[Gallery](#) / [DAX Weather Project](#) / 

[< Back](#)  
**DAX Weather Project**

Tags

EnvironmentTransportation

Required Services

0

Modified

May 22, 2020

This project includes the NOAA Weather Dataset - JFK Airport (New York) from the Data Asset Exchange and supporting notebooks. The notebooks teach the user to extract, clean and analyze sample weather data and predict weather trends to help airports schedule better flight times. This sample project contains 3 notebooks and 1 CSV file. Please run the notebooks in sequential order of their part numbers using a Python 3.6 runtime.


Images

Assets

Info



# Access from Cloud Pak for Data



Search Cloud Pak for Data product:

Cloud Pak for Data product hubHomeResourcesDocsAPIsSupportWhat's newCommunityGet support

Table of contents

Version 3.0.1 (latest)

Overview

Use cases

Planning

Installing

Services and integrations

Services in the catalog

Services outside the catalog

External data sets

Industry accelerators

Integrations

Administering

Analytics projects

Accessing data

Governing and curating data

Integrating and preparing data

Analyzing data

AI solutions

Developer resources

Troubleshooting

IBM Cloud Pak for Data > Services and integrations >

## External data sets

Transactional data is central to your business. But if you only analyze your internal transactional data, your analysis is incomplete. Complement your data with external data sets that give you a 360-degree view of your business landscape. The data in the external data sets can help you complete a more comprehensive analysis that can help you make better decisions.

IBM® Cloud Pak for Data has partnered with industry leaders to provide easy and seamless access to external reference data that you can use to enrich your transactional data. Some of the data sets provide historical data, while others provide real-time data.

The data sets make it easy for data scientists to access the data that they need from the same platform where they run build and run their analytic models.

Data offering	Provided by	Pricing	Learn more
Weather Company Data Limited Edition	The Weather Company*	Included with Cloud Pak for Data	<p><b>About this offering</b></p> <p>90-day access to cloud-based APIs that enable you to obtain historical weather data, current conditions, and forecast conditions.</p> <p><b>Use cases</b></p> <p>You can use weather data to optimize operations, reduce overhead costs, increase safety, and uncover new revenue opportunities. For example, you can:</p> <ul style="list-style-type: none"><li>Predict power outages with greater accuracy so that you can restore power to customers faster</li><li>Reduce utility costs with smarter vegetation management</li><li>Improve flight safety, efficiency and performance</li><li>Keep policyholders safe while reducing insurance claims and fraud</li><li>Improve supply chain visibility and minimize weather-related disruptions</li><li>Transport people and goods more safely</li></ul> <p><b>Industry accelerators</b></p> <p>The following industry accelerators can help you get started with this data set:</p> <ul style="list-style-type: none"><li><a href="#">Manufacturing Analytics with Weather</a></li><li><a href="#">Retail Predictive Analytics with Weather</a></li><li><a href="#">Sales Prediction using The Weather Company Data</a></li></ul> <p><b>Get started</b></p> <p>For details, see <a href="https://www.ibm.com/weather">https://www.ibm.com/weather</a>.</p>

[https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ\\_current/svc-nav/data-sets.html](https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_current/svc-nav/data-sets.html)

# Industry Accelerator - Cloud Pak for Data

## Cloud Pak for Data

View Only

Group Home

Blogs 0

Members 3

### Effective Farming - Monitor Crop Growth

0 Recommend

28 days ago

The accelerator is created using Data Asset eXchange data to support effective farming by monitoring crop growth using crop guide and provide timely alert to farmers about weather change, possible development of crop disease, evaporation of fungicide, and efficient use of solar panels (agrivoltaics support).

#### What's included?

- A structured business glossary of 90 business terms.
- Sample data science assets

#### How does it work?

The glossary provides the information architecture that you need to understand weather related business measures. Your data scientists can use the sample notebooks, predictive models and dashboards to accelerate data preparation, machine learning modeling, and data reporting. Moreover, the data scientists may modify the sample notebooks for other business use cases and corresponding datasets.

Timely alert to farmers can save crop life and bring in more cost savings.

When you import the accelerator:

- The terms are added to your business glossary under the Effective Farming - Monitor Crop Growth category in the Industry Accelerators category.
- The data science assets are added to a new analytics project.

#### Statistics

0 Favorited

17 Views

0 Files

0 Shares

0 Downloads

[https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ\\_latest/svc-nav/head/industry-accel-svc.html](https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_latest/svc-nav/head/industry-accel-svc.html)

# Hands-on tutorial

- Learn how to create and run a notebook pipeline
  - Create pipeline using Elyra Visual Pipeline Editor
  - Run pipeline locally in JupyterLab
- Explore the Data Asset Exchange
- Estimated time: 45 minutes
- Requires local Docker installation (recommended) or web browser to run in sandbox environment
- Open <https://github.com/CODAIT/DDC-data-and-ai-2021-automate-using-open-source> in a web browser and follow the instructions

# Thank You, and next steps

- **Learn more about Elyra**
  - [Latest release information](#) (new features, articles, etc)
  - [Community](#), [GitHub repository](#), [documentation](#)
- **Learn more about the Data Asset Exchange**
  - [Home page](#)
  - [Getting started with the Data Asset eXchange](#)
  - [Open data license information](#)