

DSA 8430

Parallel Computing for Data Analytics

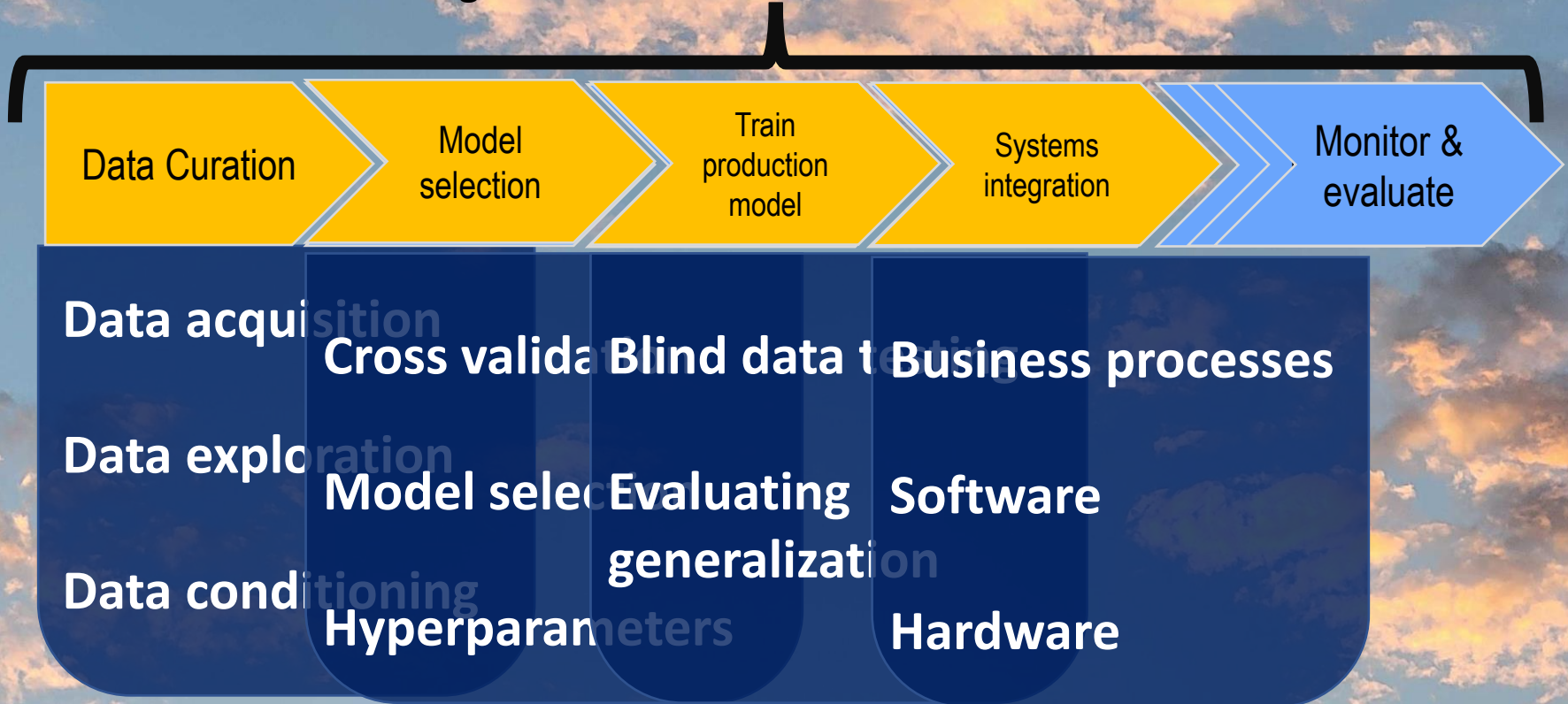
AWS SageMaker

Module Topics

- Data Science Pipelines
- SageMaker
- SageMaker Studio
- SageMaker Debugger
- Module 7 Activities
- Final Project (Module 8) – Co-Released

Data Science Pipelines

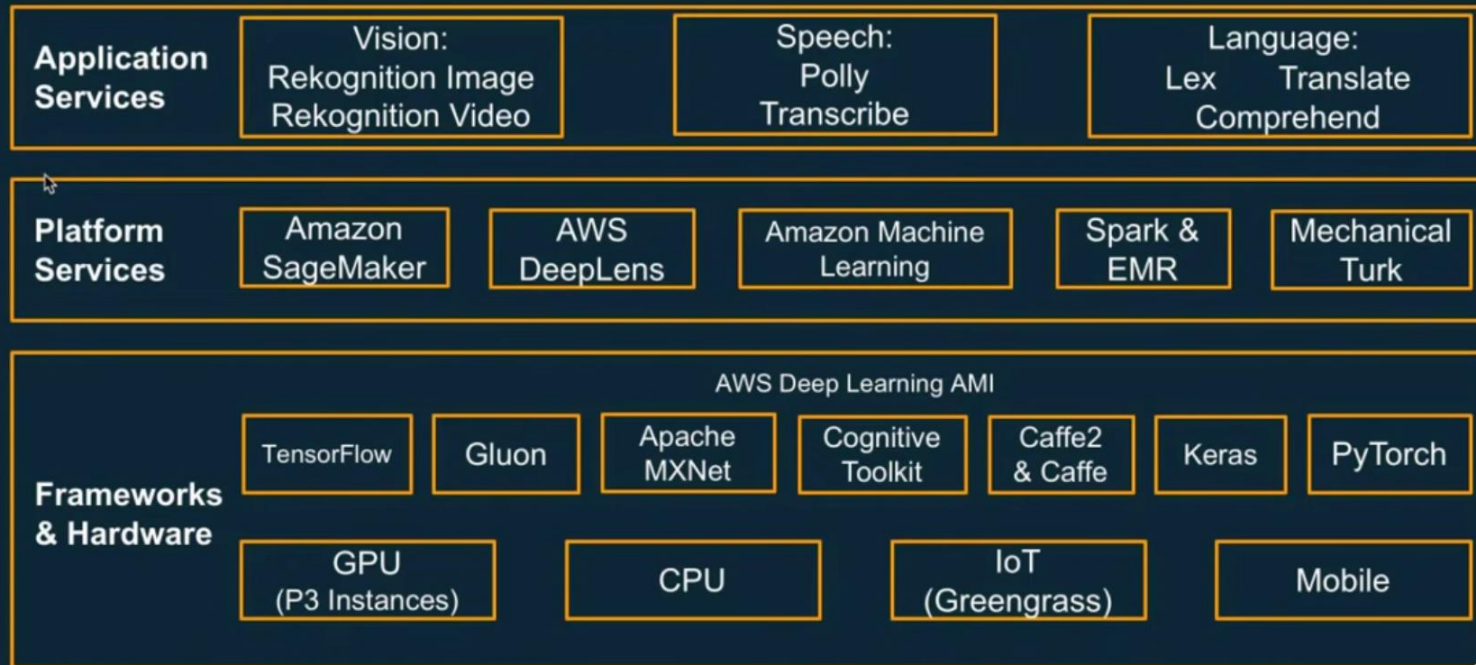
AWS SageMaker seeks to automate this!



From IEEE “Machine Learning Algorithms, Models, and Systems Integration” AWS SageMaker

SageMaker

AWS ML & AI



SageMaker Studio

The screenshot displays the Amazon SageMaker Studio web interface in a browser. The address bar shows the URL: `d-yurvtxnzau5w.studio.us-east-1.sagemaker.aws/jupyter/default/lab?`. The interface has a dark theme and includes a top navigation bar with menus for File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. On the left is a sidebar with a file explorer showing a directory structure with columns for Name and Last Modified. The main content area is titled 'Launcher' and contains two primary sections: 'Get started' and 'ML tasks and components'. The 'Get started' section features two cards: 'JumpStart models, algorithms, and solutions' with links to SageMaker JumpStart solutions like 'Detect malicious users and transactions' and 'Demand forecasting', and 'Build models automatically' with links to SageMaker Autopilot, including a video and a blog post. The 'ML tasks and components' section has four cards: 'New compilation job', 'New feature group', 'New data flow', and 'New project', each with a brief description and a '+' icon to initiate the task. The bottom status bar shows 'Git: refreshing...' and a 'Launcher' label on the right.

SageMaker Studio Clone Examples

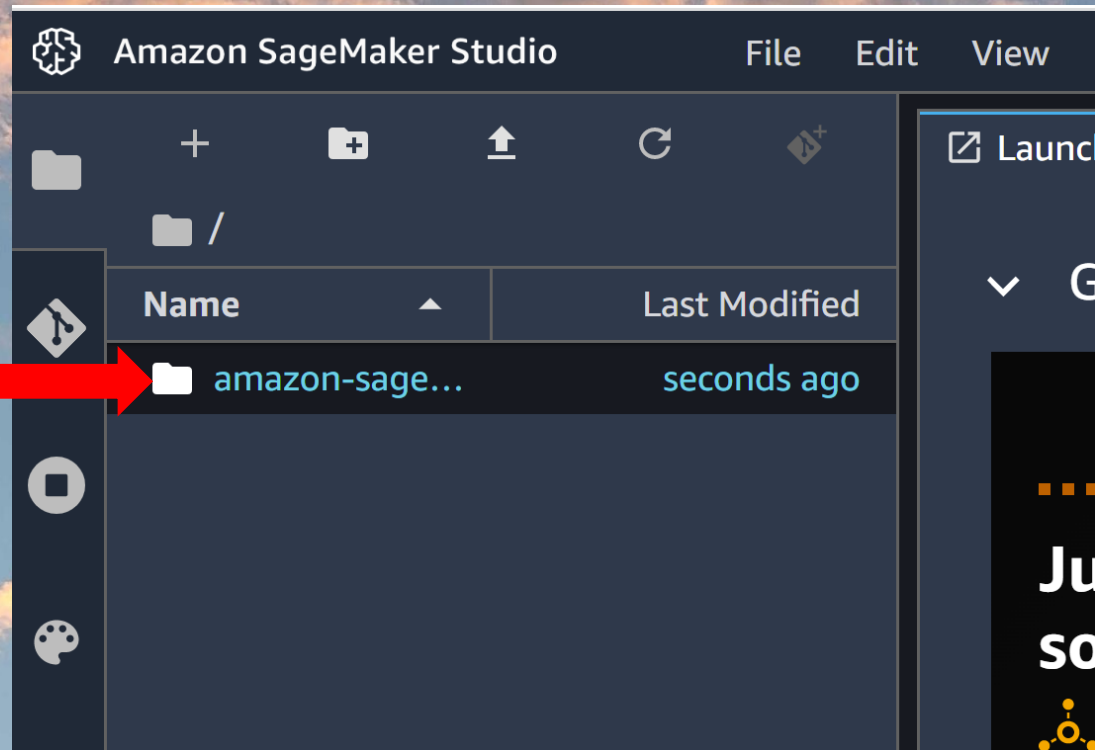
The screenshot shows the Amazon SageMaker Studio interface. The top menu bar includes File, Edit, View, Run, Kernel, Git, and Tabs. The main workspace displays a message: "You are not currently in a Git repository. To use Git, navigate to a local repository, initialize a repository here, or clone an existing repository." Below this message are three buttons: "Open the FileBrowser", "Initialize a Repository", and "Clone a Repository". The "Clone a Repository" button is highlighted with a red rectangle. To the right, a "Launcher" panel shows "Get started" options, including "JumpStart solutions".

A "Clone a repo" dialog box is open, prompting the user to "Enter the Clone URI of the repository". The URI "https://github.com/aws/amazon-sagemaker-examples.git" is entered in the text field. The dialog box has "Cancel" and "CLONE" buttons.

At the bottom of the interface, the status bar shows the Git icon and the text "Git: cloning repository..."

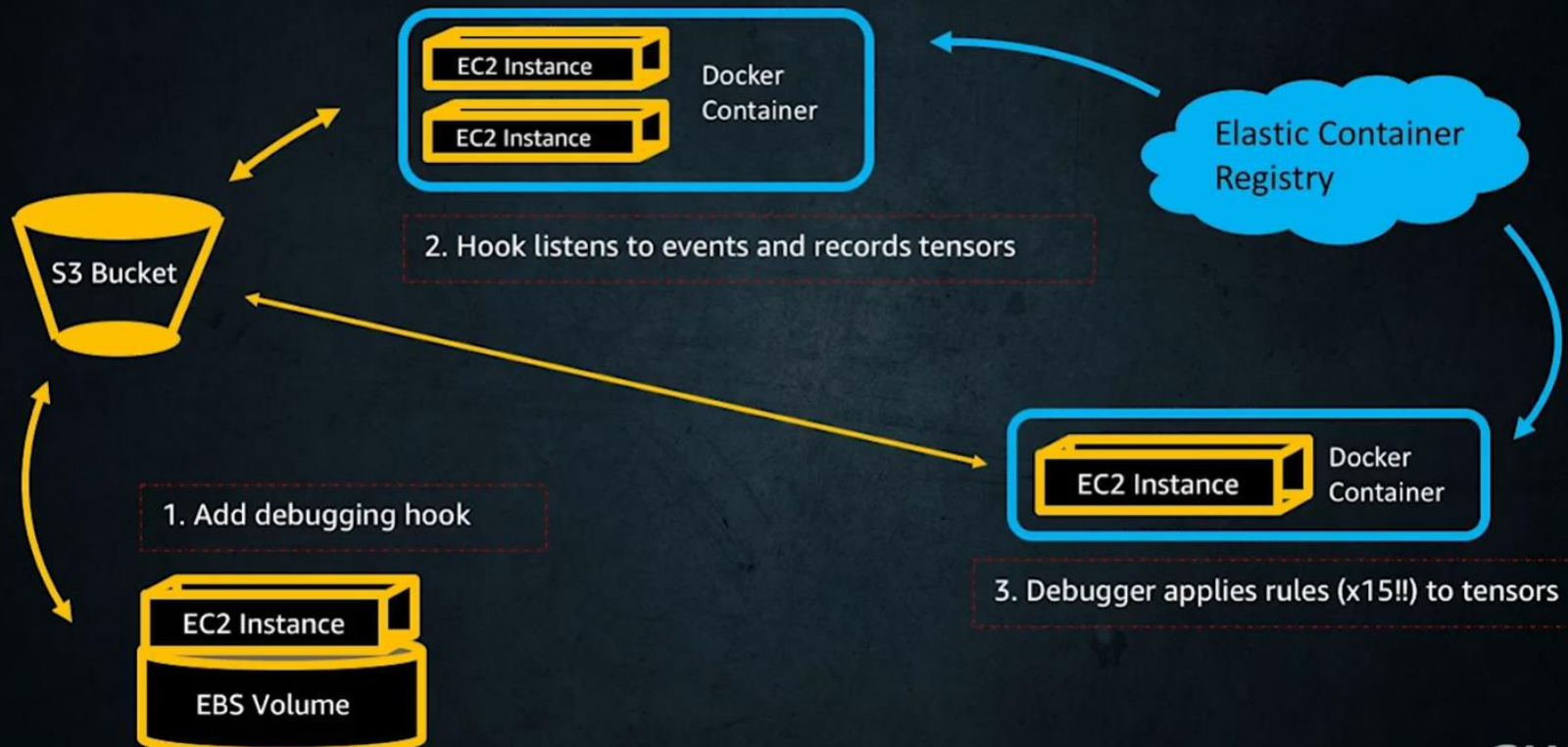
SageMaker Studio Clone Examples

Clone
accomplishes
same as “git
clone” operation
you did on
Nautilus



SageMaker Debugger

Add SageMaker Debugger Hooks to Training Jobs



SageMaker Data Wrangler

Import Data Flow Get help


Data flow

Choose the plus sign to add a step to the flow. Select a step to modify.

Create job


Validation complete 0 errors Done

Source - sampled




S3: titanic3.csv

Data types



Transform: titanic3.csv

Steps (5)



+

-

↺

↻

Look Ahead – Module 7

Module 7 will cover tutorial topics such as

- SageMaker Notebooks
- SageMaker Studio
- SageMaker Data Wrangler
- Data Bias and ML Fairness
- Building from-scratch notebooks, exports, and other downloads to provide artifacts.
 - Artifacts for **both Labs and Practices**

Look Ahead – Module 8 / Final Prj

- **No exercise for Module 7**
- Instead, concurrent release of Module 8
- **M8 = Final Project to build a SageMaker project**
 - Data Import
 - Data Wrangler
 - Data Bias Exploration
 - ML Model Training
 - ML Fairness Measurement
- **DO NOT USE SageMaker Autopilot**

Pay Attention to Detail!

This is a fitting final technology for the course because it is very complex and a lot of concepts are leveraged for distributed and parallel data analytics

Start thinking of your final project ideas as you work through M7