A. What is AWS Sagemaker?

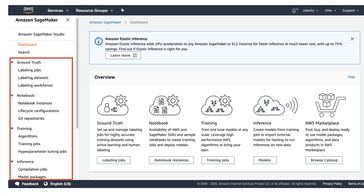
AWS (or Amazon) SageMaker is a *fully managed* service that provides the ability to build, train, tune, deploy, and manage large-scale machine learning (ML) models quickly. Sagemaker provides tools to make each of the following steps simpler:

1. Explore and process data
    - Retrieve
    - Clean and explore
    - Prepare and transform
2. Modeling
    - Develop and train the model
    - Validate and evaluate the model
3. Deployment
    - Deploy to production
    - Monitor, and update model & data

The Amazon Sagemaker provides the following tools:

- Ground Truth - To label the jobs, datasets, and workforces
- Notebook - To create Jupyter notebook instances, configure the lifecycle of the notebooks, and attache Git repositories
- Training - To choose an ML algorithm, define the training jobs, and tune the hyperparameter
- Inference - To compile and configure the trained models, and endpoints for deployments

The snapshot of the *Sagemaker Dashboard* below shows the tools mentioned above.



> *IMPORTANT NOTICE: This is the current AWS UI as of April 6th, 2020. The AWS UI is subject to change on a regular basis. We advise students to refer to AWS documentation for the above process.*

A.1. Why is SageMaker a "fully managed" service?

SageMaker helps to reduce the complexity of building, training and deploying your ML models by offering all these steps on a single platform. SageMaker supports building the ML models with *modularity*, which means you can reuse a model that you have already built earlier in other projects.

A.2. SageMaker Instances - Important to Read

SageMaker instances are the dedicated VMs that are optimized to fit different machine learning (ML) use cases. The supported instance types, names, and pricing in SageMaker are different than that of EC2. Refer the following links to have better insight:

- Amazon SageMaker ML Instance Types - ***See that an instance type is characterized by a combination of CPU, memory, GPU, GPU memory, and networking capacity.***
- Amazon EC2 Instance Types - To have you know the difference in naming and combinations of CPU, memory, storage, and networking capacity.

A.3. Supported Instance Types and Availability Zones

Amazon SageMaker offers a variety of instance types. Interestingly, ***the type of SageMaker instances that are supported varies with AWS Regions and Availability Zones***.

- First, you need to check the List of the AWS Regions that support Amazon SageMaker.
- Next, you can check the various available Amazon SageMaker ML Instance Types, again.

A.4. Instances Required for Deep Learning

The table below describes the three types of SageMaker instances that you would use in this course:

SageMaker Instance	vCPU	GPU	Mem (GiB)	GPU Mem (GiB)	Netw Perforr
`ml.t2.medium`	2	-	4	-	Low Mode
`ml.m4.xlarge`	4	-	16	-	Hig
`ml.p2.xlarge`	4	1xK80	61	12	Hig

In this course, the `ml.m4.xlarge` is needed at an early stage, while `ml.p2.xlarge` is needed only when working on the for the first project: Deploying a Sentiment Analysis Model.

> ***Note***
> *Sagemaker quotas, also referred to as limits, are very tricky. Every AWS user does not get the default quotas for SageMaker instances, which is why the last column shows a range, e.g., 0 - 20. The **Default Quota** depends on the instance type, the task you want to run (see table above), and also the region in which the Sagemaker service is requested. Refer this document having a caveat that new accounts may not always get the default limits.*

Recommended Read

AWS Sagemaker FAQs