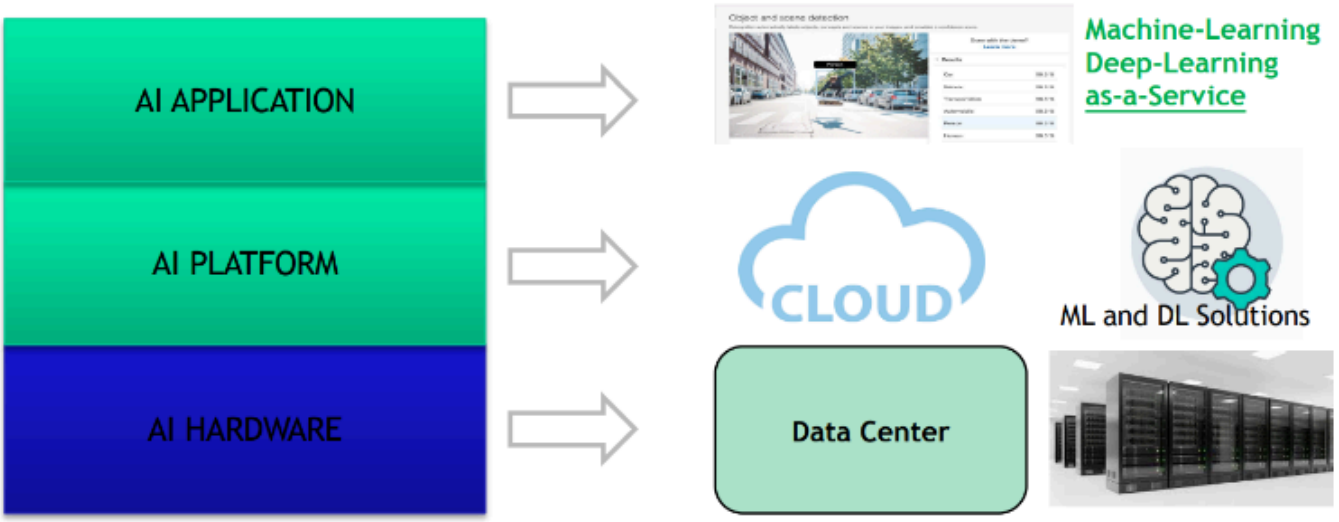
MLaaS

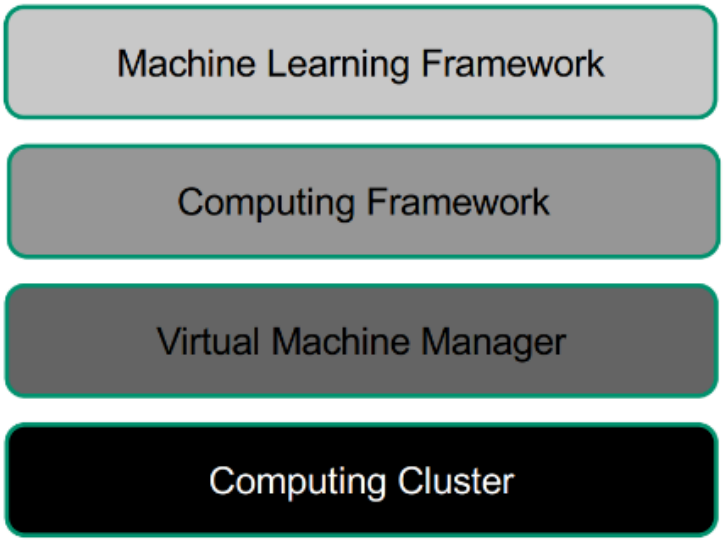
It is true that generative AI was unexpected. Even technological forecaster could't see the coming of generative AI.
It changes the availability of the algorithm as it becomes available in a browser.
It becomes a service available for the public and for the user is only a service.
This model is for inference and training. It is just for the user.
The generative AI is in the peak of information perspective. It will arrive at a plateau when it becomes part of our lives.
Now take the technologies in consideration. It was tried to give the Turing test to the chatbot and was discovered that chatgpt-4 wasn't considered as a human as it has a given dictionary and its answering have been too precise/polite in respect of human response. A chatbot developed in the '60 was recognised for a lot of cases as a human as it was rude and incorrect enough.
Generative AI can help in generating cases not included in the model training.

Cloud and ML/DL as-a-service



The HW is a datacenter with associated HW for ML. Going up we have the cloud as we are using its approach for our application. Then the application that is running is based on ML or DL.

The IT architecture for ML/DL on Datacenters



Start from the Computing Cluster(physical resources), Virtual Machine Manager where we virtualize the resources even if it degrade the performance in access, then we have the Computing Framework where we support large scale computation and then we have the layer for training.

Computing Cluster

We have servers that can be parallelized and scaled to improve performance, we have HW accelerators(CPUs and GPUs/TPUs). Storage as we need multiple storage as DAS/NAS/SAN to store the huge dataset used for training our models. Then in regards of the network it is designed to support an high workflow and not inference. The time in training is much larger than the one for inference.

Virtual Machine Manager

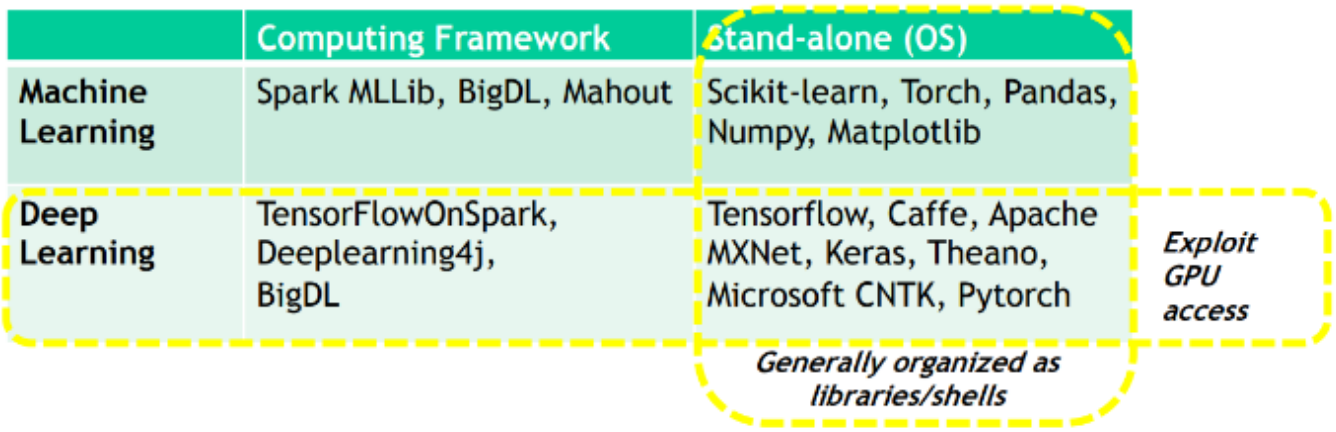
We need virtualization to give more resources to our models in the way of more VM(Many users want to access at the same time the site of the generative AI). We need platforms and resources to support virtualization

Computing Framework

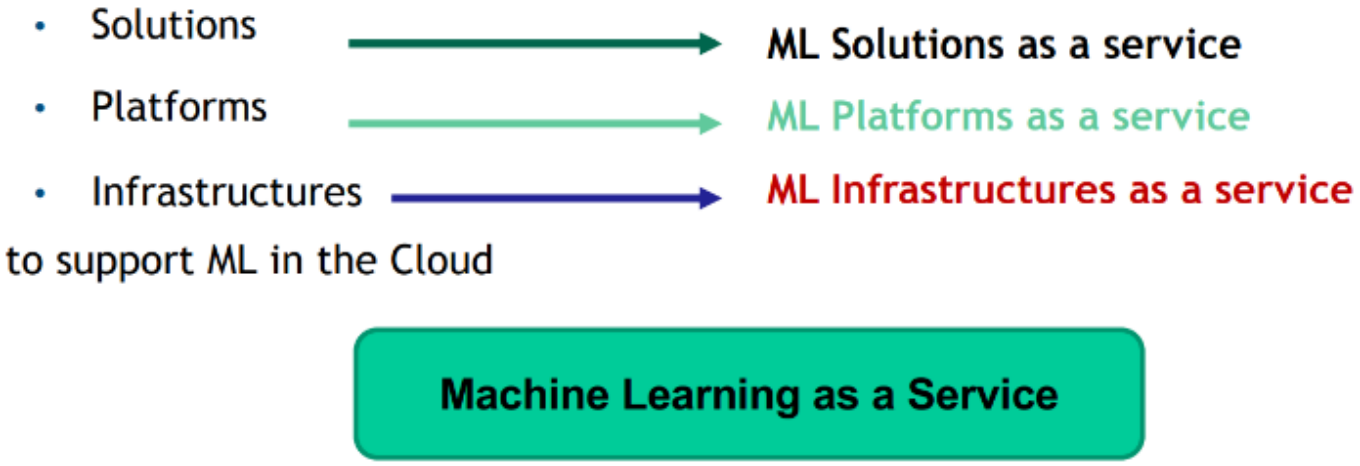
I can create a superpowerful VM but scaling it up will be expensive/time consuming so I can exploit multiple less powerful machine joint together. I can parallelize the computation on multiple machine scaling up my work. Use Cluster and Cluster Manager for parallelization.

Machine/Deep Learning Frameworks

Machine learning frameworks cover a variety of learning methods for classification, regression, clustering, anomaly detection, and data preparation, and it may or may not include neural network methods.
Deep learning frameworks cover a variety of neural network topologies with many hidden layers.

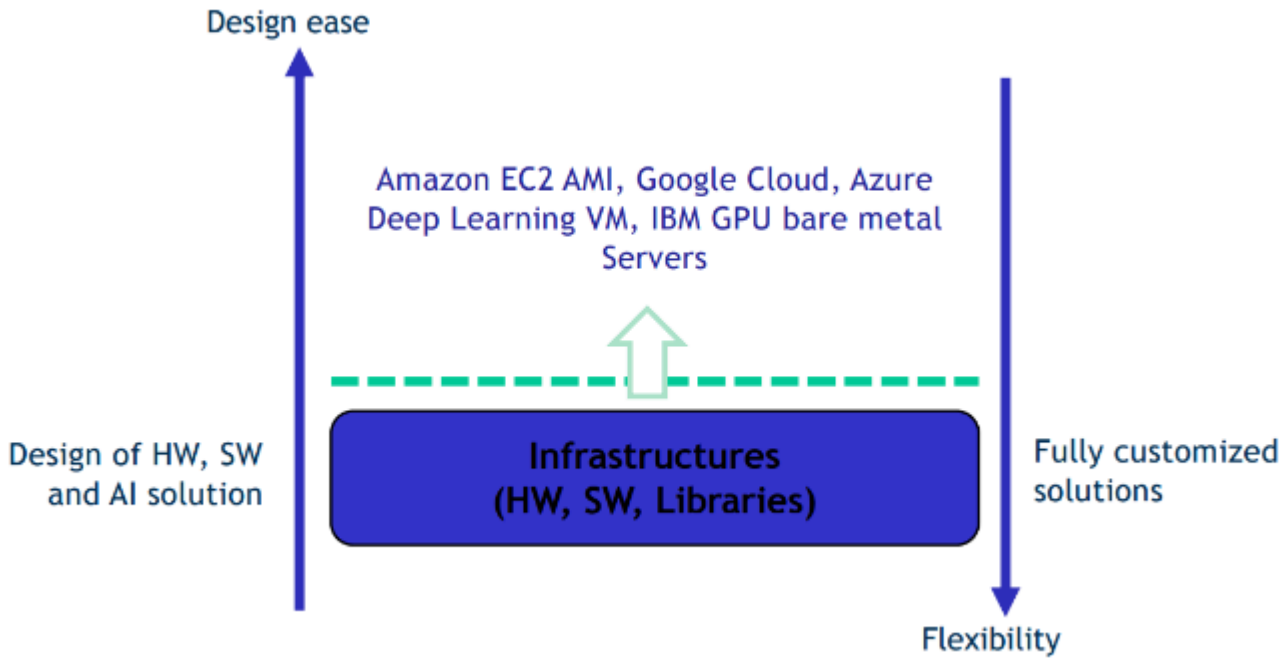


Usually the models can natively execute HW acceleration.
We can use containers to distribute our code in a way we can be sure that the code works as we want. Cloud Computing simplifies the access to ML capabilities for designing solutions and setting up a project.



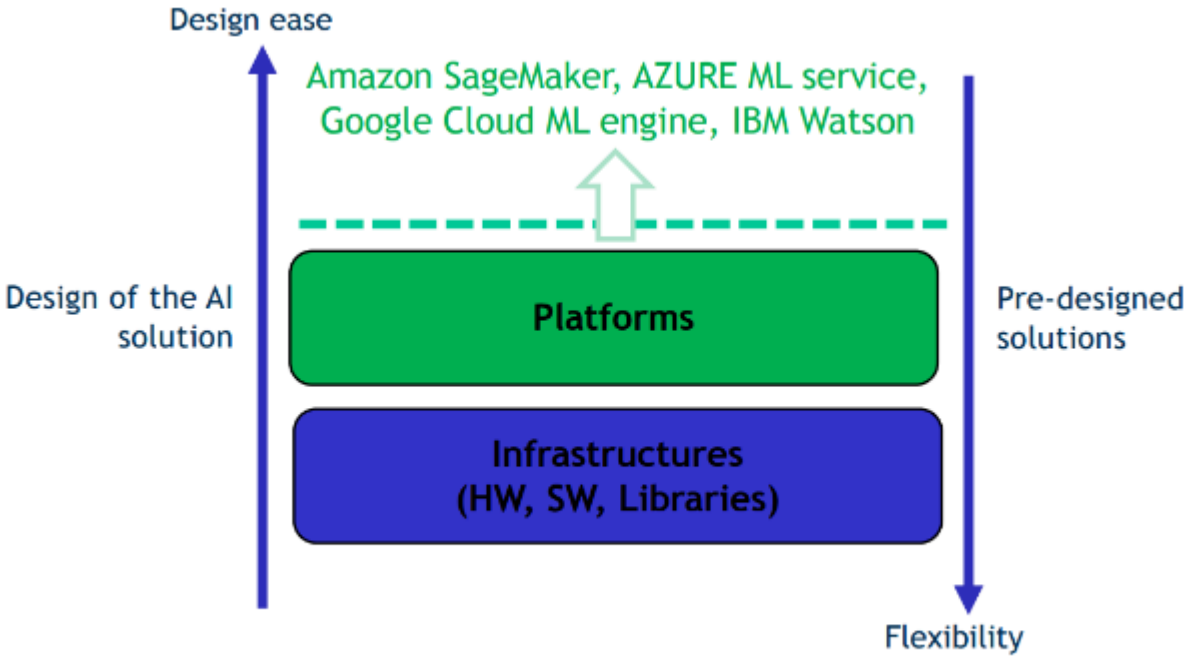
When we discuss MLaaS it take into account these three layers.

Machine Learning Infrastructure as a Service



Set of solutions available, we need to choose the best one for us. These solutions support HW for AI. I need HW, SW and solution design. This level support the HW. Higher level of design but it will be fully customizable.

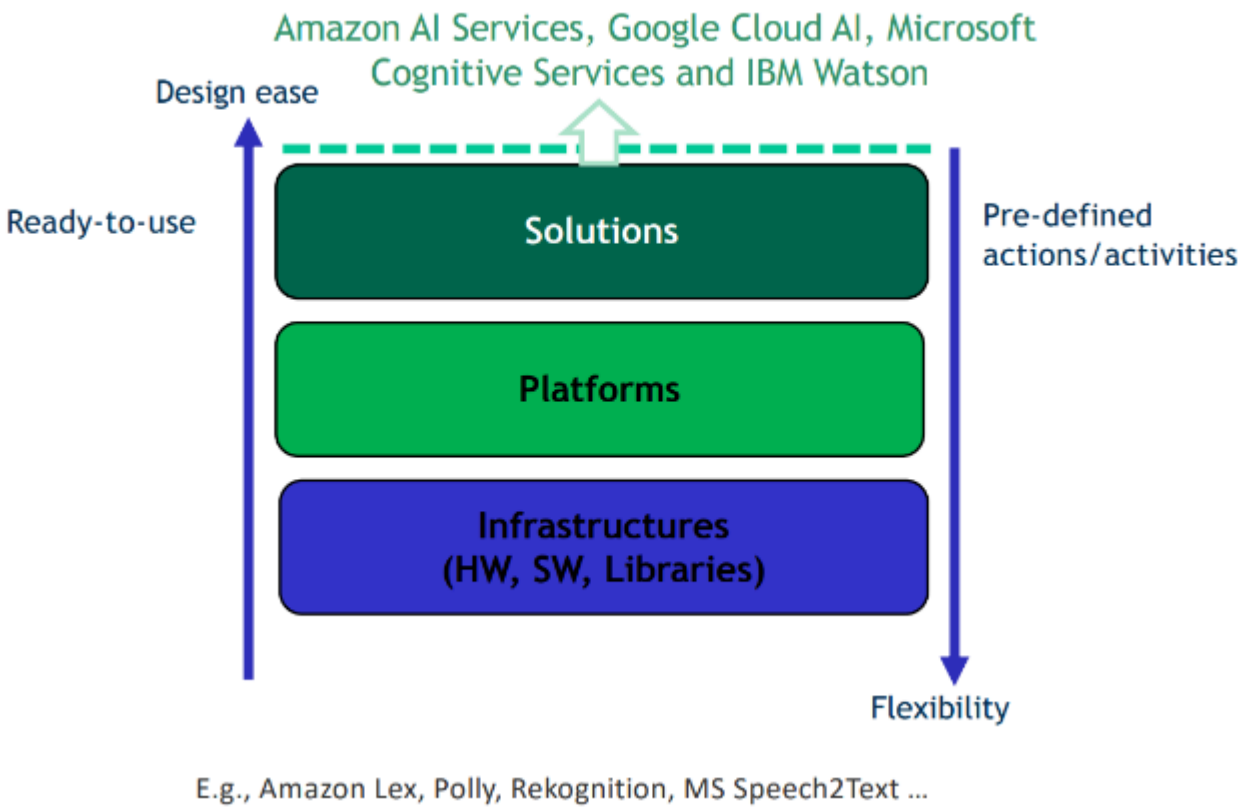
Machine Learning Platform as a Service



Provide pre-configured environments used by AI experts to train, tune and host models

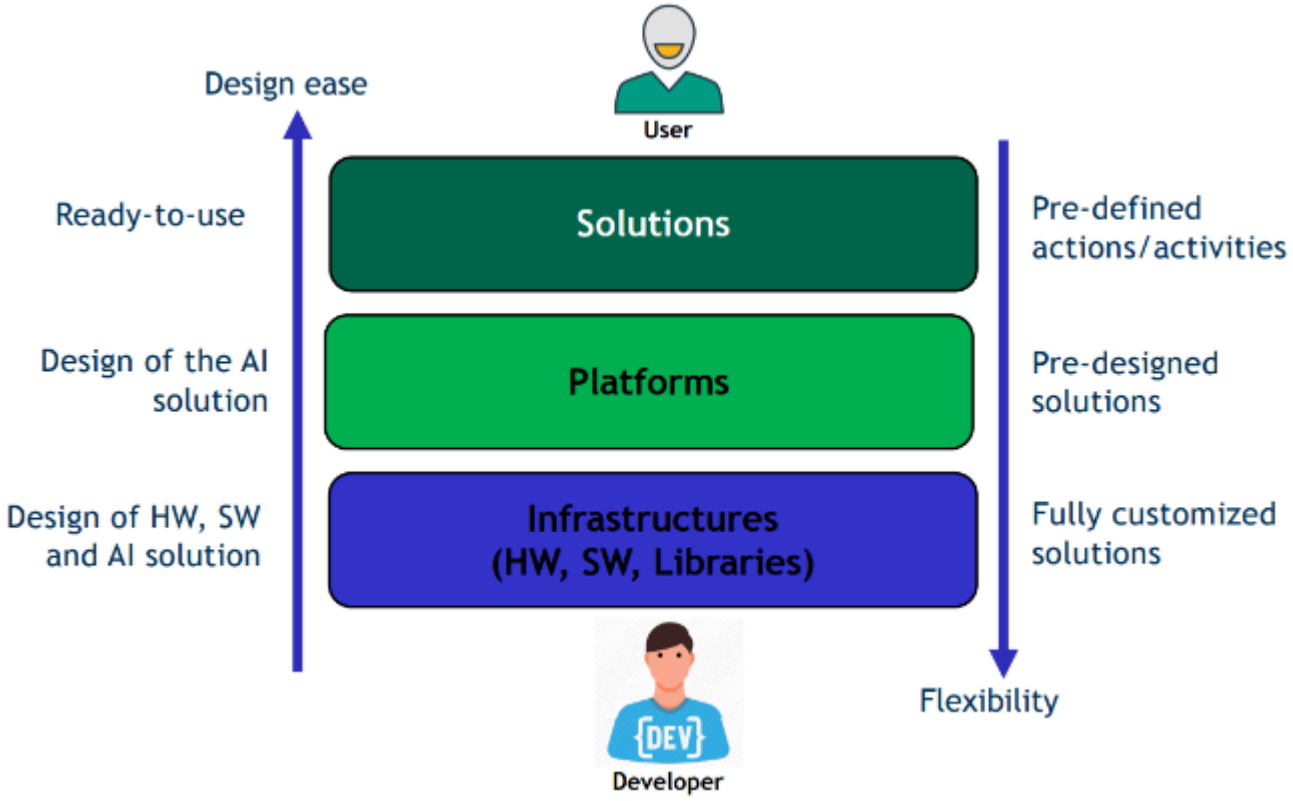
Focus on AI. You don't worry about HW, it is taken care by the platform. You only have to write the application. You give instruction about what you want to do and don't talk about HW.

ML software as a service



Someone else programmed, trained and deployed the model you only have to use it. A lot of services. These solutions weren't invented as a service. The constraint is on the set of classes that can be used as they are defined by someone else.

AI and off-the-shelf technological solutions



This is a fully customizable solution. You can choose between some solution. At Solutions level there are some solution already usable but they are not customizable.
You need to keep in mind that using external services put you in the hands of the service providers as they can change/discontinue the service. You have to make the Make or Buy decision.