Machine Learning in 15: Getting started with Generative AI using Amazon SageMaker

Philipp Schmid

Technical Lead
Hugging Face

Heiko Hotz

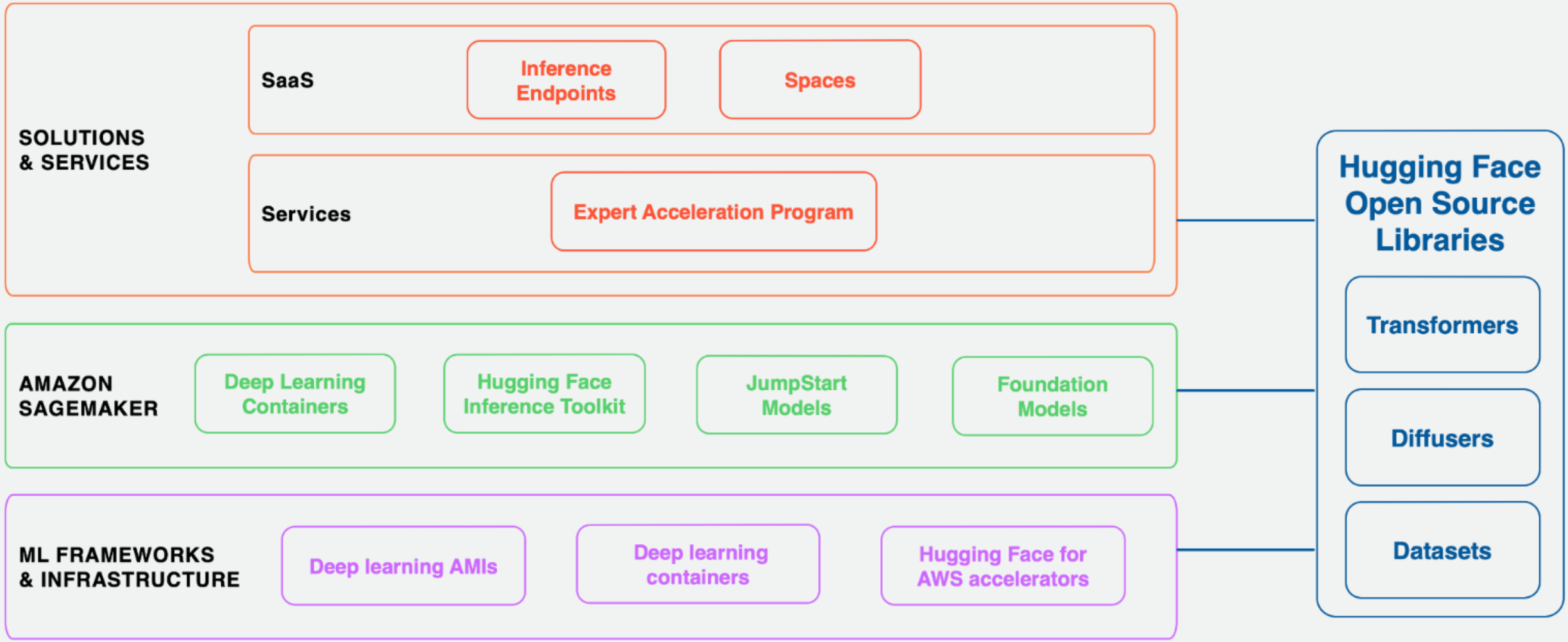
Solutions Architect for AI/ML
AWS

What is Hugging Face

Hugging Face is the largest open platform for Artificial Intelligence (AI) and Machine Learning (ML), called the "Hub". Initiated in 2019, it has been called the "Github of machine learning" thanks to the power of open science and open source.

The Hugging Face Hub hosts over 200,000 pre-trained open-source models and over 30,000 open-source datasets for many ML task types (text, images, video, audio). Customers can fine-tune these models on proprietary data for increased relevance and accuracy instead of using unpredictable closed LLMs. Building their own models also helps customers protect their intellectual property and level up their capabilities.

Hugging Face on AWS

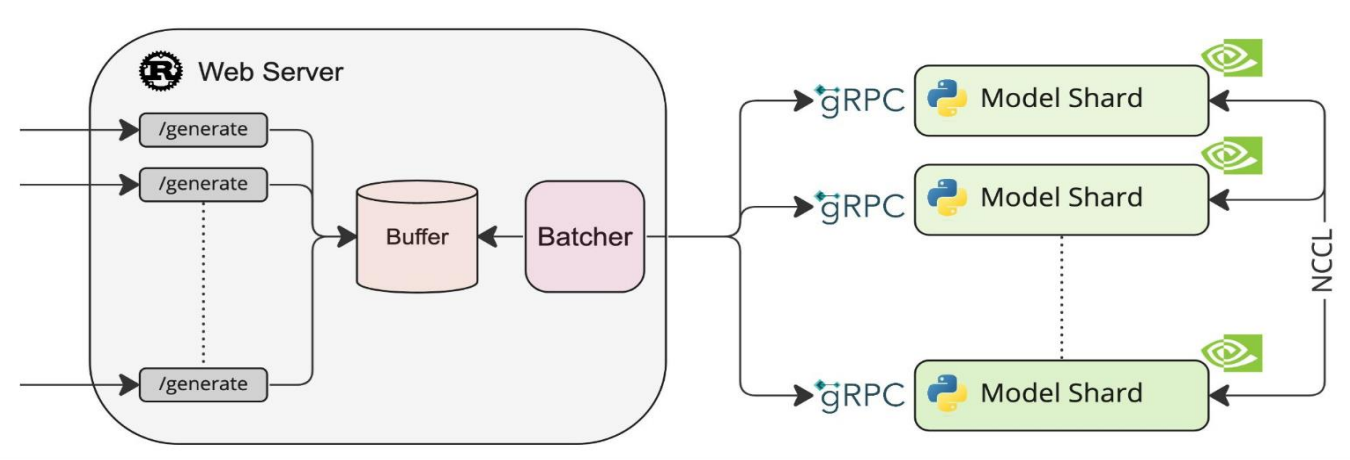


Customer pain points when running LLMs:

- **Computational Resources:** LLMs require a lot of memory and computational power to run effectively
- **Hardware Requirements**: LLMs are often too big to fit into a single machine and require complex parallelism and communication, which creates challenges for deployment (I/O) and scaling.
- **Latency & Throughput:** The Hardware requirements oftentimes have a negative impact on Latency & throughput, which requires optimizing models, as efficiently leveraging compute is critical.
- **Maintenance:** LLMs are complex systems and require ongoing maintenance and monitoring.
- **Security & Governance:** Many LLMs are currently not available in open source and only usable through commercial APIs, which makes it hard or impossible for enterprises to leverage them since it is not possible to use them without exposing customer or confidential information or support governance and control processes regarding training data, and IT systems logging.

Hugging Face LLM Inference Container for Amazon SageMaker

Hugging Face LLM DLC is a new purpose-built Inference Container to easily deploy LLMs in a secure and managed environment. The DLC is powered by Text Generation Inference (TGI), an open-source, purpose-built solution for deploying and serving Large Language Models (LLMs).



Hugging Face LLM Inference Container for Amazon SageMaker

- Natively integrate the container into Amazon SageMaker platform
- Solve challenges for running LLMs on Amazon SageMaker
- Text Generation Inference is a solution by Hugging Face for optimized inference on open source LLMs
- Powering Hugging Face services at scale (BLOOM, GPT-NeoX, StarCoder, LLaMA, Falcon, T5 etc.)
- Used by IBM, Grammarly, and Open-Assistant initiative

Hugging Face LLM Inference Container for Amazon SageMaker

Deploy Falcon 40B

```python
llm = HuggingFaceModel(
  image_uri=get_huggingface_llm_image_uri("huggingface",version="0.8.2"),
  env= {'HF_MODEL_ID':'tiiuae/falcon-40b-instruct','SM_NUM_GPUS': '4'},
  role=role,
).deploy(...)

# send request
llm.predict({
  "inputs": "What is Amazon SageMaker?",
})
# Amazon SageMaker is a fully managed...
```