Chapter 9: Generative AI on AWS

In the previous chapters we focused on running generative AI models locally: your own laptop, a beefy desktop, perhaps a Kaggle notebook or a Colab – the last two are not, strictly speaking, “local” but all those options share a certain similarity. Be it through scaling needs or inference / deployment needs, at some point you will head towards the cloud (a.k.a. “somebody else’s computer”) and in this chapter we take the first of three steps on this part of the journey.

# Learning path

While deep learning techniques have been extremely successful in computer vision tasks like image recognition, segmentation, or object detection, the underlying algorithms are typically extremely data intensive: they require large amounts of data to avoid overfitting. Not all domains of interest satisfy that

**A picture containing text, screenshot, diagram, design

Description automatically generated**

Figure 9.1: Landing page for the Generative AI track from GCP

# Generative AI Studio

Asd

# Running models on GCP

Asd