Cost Analysis for Azure

Breda University of Applied Sciences

26th June, 2025

Natural Language Processing

Group 6

Kosta Ljubisavljevic (233101)

Ricardo Rino De Sousa (235038)

Soheil Mohammadpour (231754)

Noah Ivanisevic (235738)

Erfan Salour (230499)

Introduction and Azure Services Used

During the Natural Language Processing (NLP) project, we had to use different services from Azure. These services need to be known so we can perform a correct cost analysis. The following pipeline shows what we used:

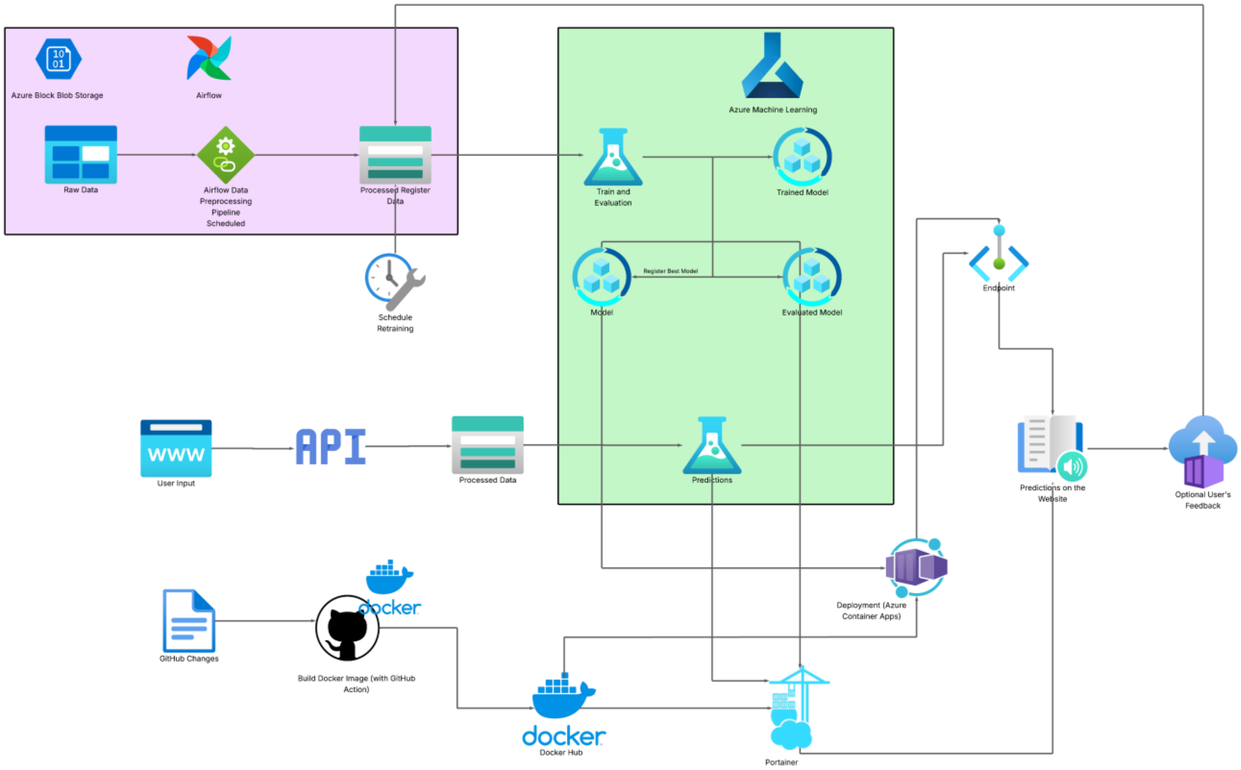


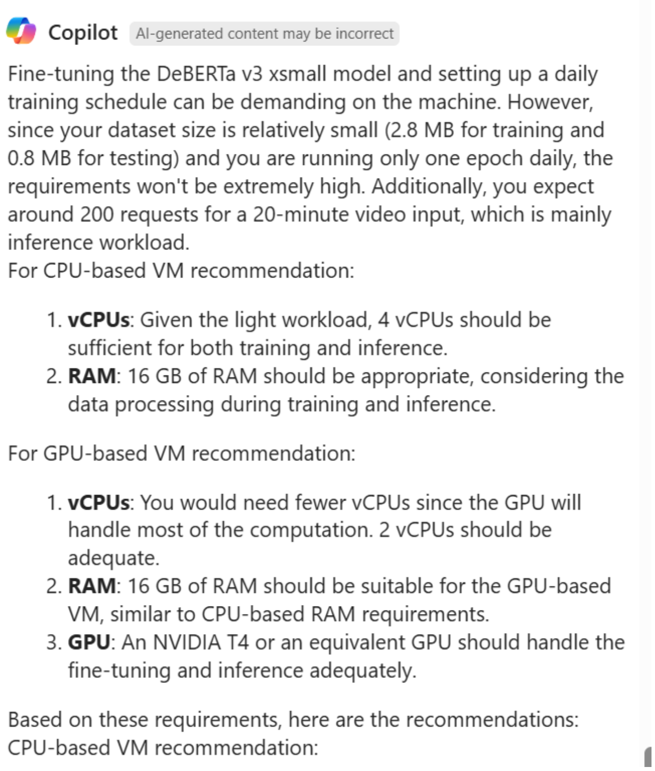
Figure 1 – Azure ML pipeline with User Interaction

From figure 1, we can see raw data and processed data. Both were stored in **Azure Block Blob Storage**, one of the options available. Other services used for this pipeline were **Azure Machine Learning**, which was where our group did the training, evaluation, and management of the model/models. We used **Azure Private Link** to create **private endpoints**, which allow secure connections to Azure services over a private network, ensuring all traffic remains within the Azure backbone. **Azure Container Apps** was used to deploy the trained model as a containerized application, including any supporting steps required for inference.

How much is the cost? How to calculate it? *West Europe Prices*

This is a theoretical situation to get an idea of the costs. We do not expect to have that increase of data unless we start having new users, but applying a more powerful CPU in Azure ML would be more expensive, so it is better to start with what can be used for the short and medium term.

First of all, the website used to perform these calculations was the Azure Pricing Calculator from Microsoft.For the Azure Machine Learning, we used the recommendations from the website, which consist of going to the Azure portal and giving exact information to copilot to give you some recommendations:



Copilot AI-generated content may be incorrect

Fine-tuning the DeBERTa v3 xsmall model and setting up a daily training schedule can be demanding on the machine. However, since your dataset size is relatively small (2.8 MB for training and 0.8 MB for testing) and you are running only one epoch daily, the requirements won't be extremely high. Additionally, you expect around 200 requests for a 20-minute video input, which is mainly inference workload.
For CPU-based VM recommendation:

1. **vCPUs**: Given the light workload, 4 vCPUs should be sufficient for both training and inference.
2. **RAM**: 16 GB of RAM should be appropriate, considering the data processing during training and inference.

For GPU-based VM recommendation:

1. **vCPUs**: You would need fewer vCPUs since the GPU will handle most of the computation. 2 vCPUs should be adequate.
2. **RAM**: 16 GB of RAM should be suitable for the GPU-based VM, similar to CPU-based RAM requirements.
3. **GPU**: An NVIDIA T4 or an equivalent GPU should handle the fine-tuning and inference adequately.

Based on these requirements, here are the recommendations:
CPU-based VM recommendation:

Picture 1—Asking for Help in the Azure ML portal to find the correct Virtual Machine for Azure Machine Learning

On the pricing website, we can select it like this and change whatever we need, for example, more compute time, more instances, or better specifications. The method we used was to run the hyperparameters in sequence. If we decided to run them in parallel, it would require more instances. By doing this, it would cost more since you are using more instances but would require less time to run it. In our scenario, it would take around 7 hours to run the training, which is run daily.

Name	CPU	Memory	Price/hour	Instances	Hours	Total Price
D4ds v4	4	16 GB	0.272 $	1	210	57.12 $

Table 1 – Azure Machine Learning Pricing

When the user asks for a video that is around 20 minutes long to be transcribed, it takes around 200 requests for that video. This is for Azure Container Apps. Considering that the first 2 million requests are free of charge per month, we do not expect our application to have a real cost at the moment, since to break the 2 million free requests, we would need around 10,000 videos of 20 minutes long per month. In our hypothetical situation, it is 500 videos per month.

After that, there is a need to calculate the endpoint price. For the endpoint, we are going to keep using the base of 500 files per month. In our pipeline, the only thing that is transferred from the endpoint to the pipeline is the audio file (which for a YouTube video of 20 minutes is around 8 MB—inbound data—bringing us to a total of 4 GB per month for 500 files), and the outbound data is around 0.011 GB per month since the output is a CSV file of around 23 KB. As the actual transcription and prediction workload is executed via an Azure Machine Learning job, the container app's compute time is negligible and set to zero for pricing purposes. Regarding these 500 files of 20 minutes length for Azure ML it brings us to a total of 1:30 minutes of computing per video, which corresponds to 12.5 hours per month. The cost can be seen in the following table:

Name VM	Price/ hour	Hours	Outbound data (GB)	Inbound data (GB)	Nº Files	Average Minutes	Data Price Endpoint (per GB)	Total Price
D4ds v4	0.272 $	12.5	0.011	4	500	20	0.01 $	3.45 $

Table 2 – Azure Endpoints Pricing

Lastly, to calculate the price of Azure Block Blob Storage.

Considering our current datasets and their sizes:

Dataset	Size (GB)
Train	0.0056 (raw + processed)
Test	0.0016 (raw + processed)
Total	0.0072 (raw + processed)

Table 3 – Dataset Size

The CSV files are counted as raw and processed since both of them are stored.

In the hypothetical scenario of 500 files of 20 minutes long per month (for the next 6 months), the price would be going like this:

File Type	1st Month	2nd Month	3rd Month	4th Month	5th Month	6th Month
Audio	4 GB	4 GB	4 GB	4 GB	4 GB	4 GB
CSV (from audio)	0.022 GB	0.022 GB	0.022 GB	0.022 GB	0.022 GB	0.022 GB
CSV + audio (from dataset)	0.0072 GB	4.0292 GB	8.0512 GB	12.0732 GB	16.0952 GB	20.1172 GB
Total	4.0292 GB	8.0512 GB	12.0732 GB	16.0952 GB	20.1172 GB	24.1392 GB
Price (Capacity)	0.08 $	0.17 $	0.25 $	0.33 $	0.42 $	0.50 $

Table 4 – Data Storage Block Blob Pricing for Data

In this case, the number of operations correspond to the number of files per month

The Operations and Data Transfer:

Operations	Number of operations	Price per 10000 operations	Price
Write Operations	500	0.050 $	0.01 $
List and Create Container Operations	500	0.050 $	0.01 $
Read Operations	500	0.004 $	0.01 $
Other Operations	500	0.004 $	0.01 $
Total	2000		0.04 $

Table 5 – Data Storage Block Blob Pricing for Operations

This brings us a total of:

Services	1st Month	2nd Month	3rd Month	4th Month	5th Month	6th Month
Azure Machine Learning	57.12 $	57.12 $	57.12 $	57.12 $	57.12 $	57.12 $
Azure Container Apps	Free	Free	Free	Free	Free	Free
Azure Endpoints	3.45 $	3.45 $	3.45 $	3.45 $	3.45 $	3.45 $
Data Block Blob Storage	0.12 $	0.21 $	0.29 $	0.37 $	0.46 $	0.54 $
Total	60.69 $	60.78 $	60.86 $	60.94 $	61.03 $	61.11 $

Table 6 – Pricing of Azure during 6 months

These prices might be changed in the future, so keep yourself updated with the Azure pricing calculator if you plan to use a similar plan. The cost of Azure Machine Learning, we did not expect to change that much since the size of the CSV would not be much bigger to ask for a better CPU and more compute time (theoretical situation). For Azure Container Apps, as said previously, the first 2 million requests in a month are free of charge. We expect to have 100000 requests per month (hypothetical situation). The Data Block Blob Storage price has an extensive breakdown of the pricing and the endpoint charge based on outbound and inbound data since the computing is taken care of by Azure Machine Learning. However, this price I included in Azure Endpoints, but it is based on the Azure Machine Learning Virtual Machine.