

Introduction

Introduction to Cloud Computing

What is cloud computing

This is the availability of IT resources over the internet. Here you can pay based on how much you use.

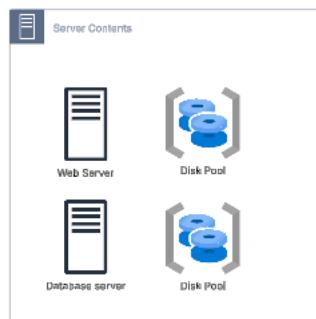
Let's say a company needs to host a web application that can be consumed by users across the world.

What does a company need to do - At a high level

1. Gather requirements
2. Design product
3. Development and Test
4. Host the application in production

Now in order to host the application in production or even have a development or test environment for the application you need to have an infrastructure in place.

We need physical or virtual servers to host the application and database layer.



We need storage to store the data.

Ethernet Switch



Then we need to build a network.
Make sure our application is reachable to users.

Network



Large companies that host a number of applications would need a vast array of servers. They would host the servers in data centers.

The company needs to invest in the data center - Rent of space, buy space, pay for cooling , electricity. They need to buy servers and storage. All capital expenditure.

And it takes time to procure hardware and set it up. Most of time you can end up with infrastructure not being utilized. Or not getting the infrastructure setup in time for hosting your application.

The screenshot shows the Amazon Q homepage with a dark background. At the top, there's a navigation bar with links for Products, Solutions, Pricing, Documentation, Learn, Partner Network, AWS Marketplace, Customer Enablement, Events, Explore More, and a search icon. Below the navigation, there are several sections: 'Featured Services' on the left listing various AWS services like Analytics, Application Integration, Artificial Intelligence, etc.; 'Featured Services' in the center listing Amazon Q, Amazon SageMaker, Amazon EC2, etc.; 'Resources and Media' on the right with links to the Blog and 'What's New on AWS'; 'Customer Enablement' with links to AWS Training and Certification, AWS Professional Services, AWS Security Assurance Services, AWS Support, AWS Managed Services, and AWS re:Post; and a sidebar on the far left listing more AWS services under 'Featured Services'.

With a cloud provider such as Amazon Web Services you get a host of web services that can be accessed via the Internet.

For a host of these services, you don't need to buy any physical hardware beforehand. You don't need to spend upfront money/capital expenditure.

For example when it comes to servers, you can host a virtual server on the cloud via the use of Amazon EC2 service. Here you don't manage the physical infrastructure , that's managed by AWS.

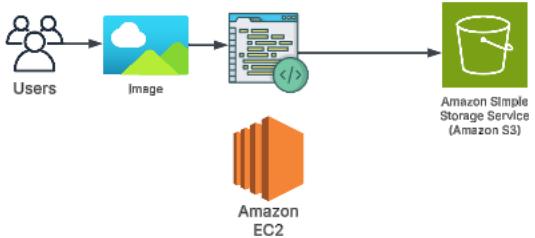
You get on-demand compute, here you only pay for how much you use. At any point in time when you don't need the compute you can terminate the compute infrastructure.

Our first AWS service - Amazon S3

Amazon S3

This is an object storage service. You can store your data as objects onto this service.

There are a lot of use cases as to where this service can be used.



The S3 service can be used to store the images.

Compute Infrastructure

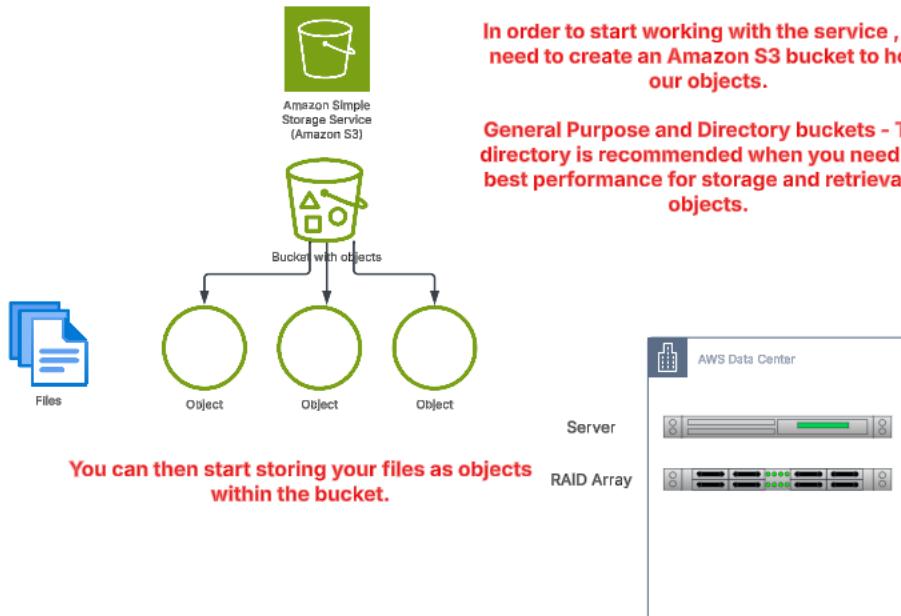
You can use Amazon S3 for data engineering - as a data lake store



Data Engineers can analyze the data and work with the data.

You can virtually store any type of file , because they are stored as binary objects.

You don't need to worry about managing the underlying storage, the physical storage is managed by the service. It is allocated by the service automatically in the background.



Remember in the end your data is stored on some sort of physical media in an AWS data center

But the physical infrastructure and storage is managed for you.



Let's work on Machine Learning

Understanding different terms

Machine Learning

This is widely used in industries today.

We can use Machine Learning models to detect fraudulent bank transactions.

We can use machine learning models to make predictions such as stock predictions.

Deep Learning



We can use Deep Learning in the field of Computer Vision. Here we can extract information from images and videos.

Natural Language Processing

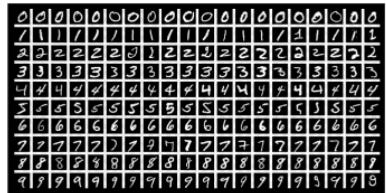


Here deep learning is used to build conversational AI solutions.

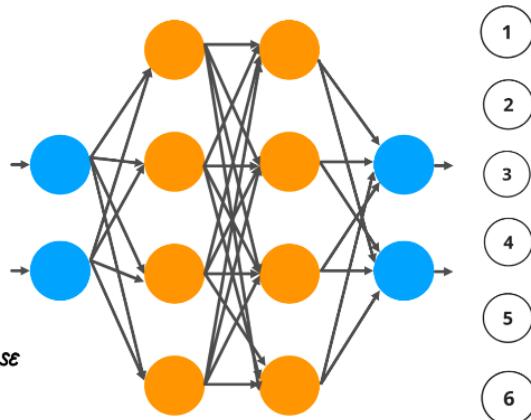
Chatbots on sites are a common example of this. These are human-like agents designed to converse with customers.

I NATURAL LANGUAGE PROCESSING

NEURAL NETWORK



REFERENCE -
https://en.wikipedia.org/wiki/MNIST_database#/media/File:MNIST_dataset_example.png



Neural networks are a type of model that is used to mimic the biological neurons of a human brain.

Here we have a representation of artificial neurons that are part of a neural network.

Considering Machine Learning



Consider an ecommerce website.

Customers purchase products from a line of products.



Developers will first use a set of programming languages to develop the web site.

**MACHINE
LEARNING**

- Would be used Machine Learning to design the interface of the web site?
- Would we use Machine Learning to code the process of a user's purchase on the site?
- Would we use Machine Learning to control the back end data?

So all of this would be done with by a development team. The code base is what makes up the application.



- What about understanding the user churn, does the user return to the site for another purchase.



- What's probably the best time to conduct a sale on the site.



- Sales forecast.



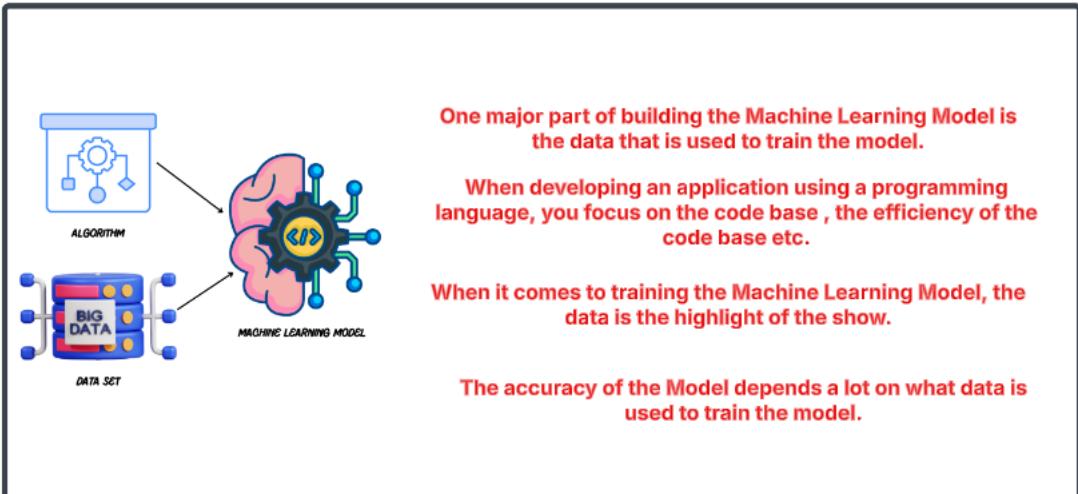
We can probably use data warehouses , data visualization tools



We can also develop Machine Learning models.

Depends on the final business requirement and what will give the maximum value add.

Data - The star of the show



One of the main reasons for the popularity of Machine Learning is the vast amounts of data that is available for training a Machine Learning Model.

Phase 1 - First identify the different data sources



Phase 2 - Make sure the data is usable

We need to feed in a rich data set to train the Machine Learning Model.

Does our data set contain missing column values for certain rows? What should we do in these cases.

Does our data set contain outlier values? How do we tackle such cases.

Do we have all data columns of Interest?



Phase 3 - Generate a feature set

Location	Number of rooms	Agent	Area (sq ft)	Price (USD)
LocationA	5	AgentA	10,000	500,000
LocationB	4	AgentB	12,000	550,000

Now from the data set , we need to extract or generate features that can be fed into our Machine Learning Model.

The Machine Learning algorithm which is mathematical in nature will look at patterns in the feature set and then use those patterns to predict future values.

Feature - This is a measurable property within the data set.

Label - This is used to signify to the Machine Learning Model , that for a set of input values, this is the output value. In the end this is the value it needs to predict. This is used for Supervised Learning.

Feature Engineering - Generate new features based on existing ones.

The features need to make sense to the Machine Learning algorithm. Make sure the features are selected on the business problem that it is being targeted to solve.

We also need to make sure we have a good spread of values in our data set that is used to train the model. There should be no bias in the data.

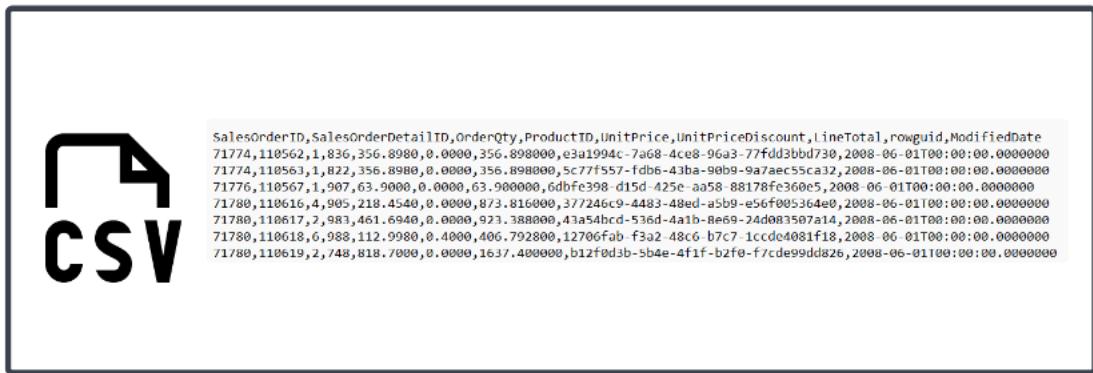
Let's say we are building a banking solution that can be used to detect fraudulent transactions.

We build a data set that has a historical set of transactions, a large data set. This data set has already been labelled. So under which conditions, parameters , features, the transaction has been detected as fraudulent.

Now if we feed in a data set that has 90% of transactions marked as non-fraudulent, the model has been trained very well on detecting non-fraudulent transactions. So even when making predictions in the real-world , it might classify actual fraudulent as non-fraudulent transactions.

Different types of data

We need to consider that data can come in different formats



The above case is an example of semi-structured data.

Another popular example of semi structured data is JSON -
JavaScript Object notation.



We can also have data in the form of tables - Structured data.

We normally use database relational engines to store structured data.



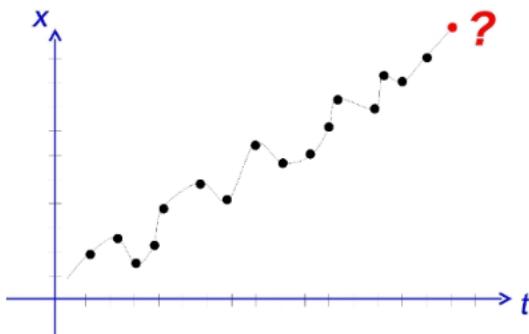
SalesOrderDetailID	OrderQty	ProductID	UnitPrice
110562	1	836	356.8980
110563	1	822	356.8980
110567	1	907	63.9000
110616	4	905	218.4540
110617	2	983	461.6940
110618	6	988	112.9980

We can use these data sets to train machine learning models for different purposes - We can use models to predict future values, classify data.

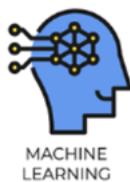
We can also get time-series data, these are data points over past time values.

We can then use these data points along with the Machine Learning Model to predict future values.

Forecasting

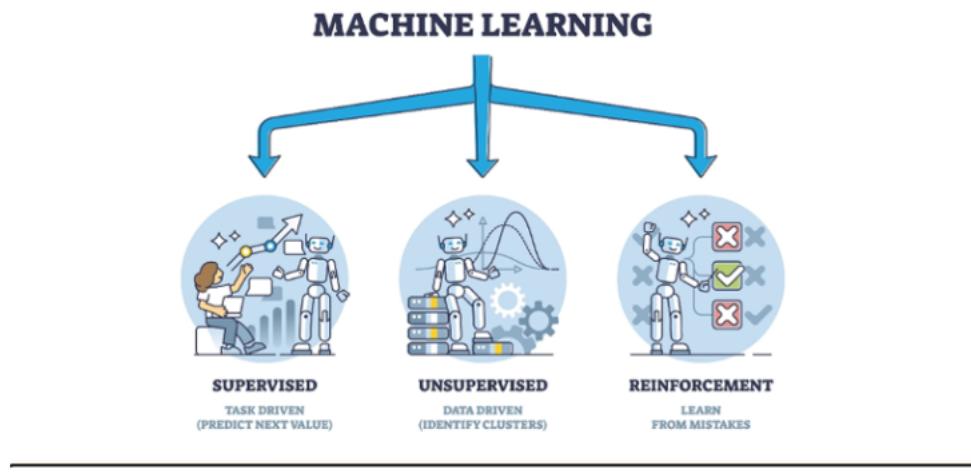


Then we have unstructured data - Here we can feed in images, audio , video to train the Machine Learning model.

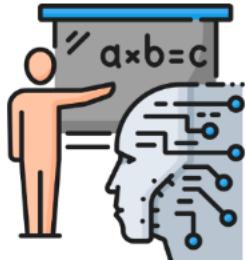


We can feed in different images to train a model to identify the picture of a cat.

Different types of Machine Learning tasks



Supervised Learning



Here we have a set of input values along with their output values when training the model. Here you have the labels for the features.

Classification



We just need to classify values.

Binary classification - Is the credit card transaction fraudulent?

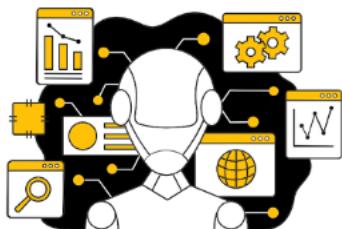
Multiclass classification - Here we want to assign the output to one of several classes

Regression



Here we want the machine learning model to forecast values.

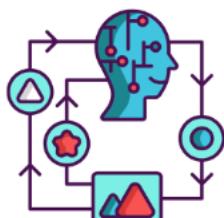
Unsupervised Learning



Here we train the model just using features.
There are no labels for the features.

This type of learning can be used to discover patterns in the data and group the data values.

Reinforcement Learning



Here the machine learning model learns through trial and error.

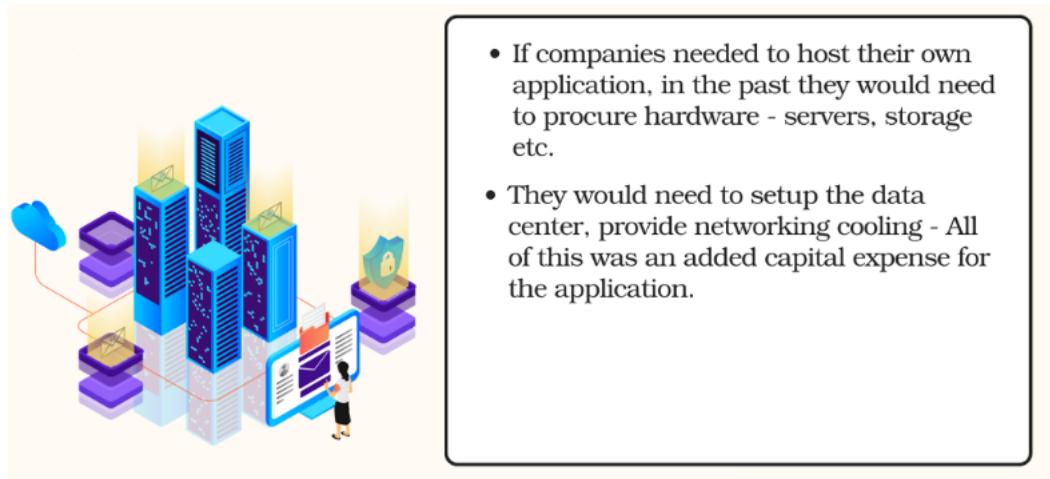
If it gets it right, it is awarded a reward so that it understands that it's on the right track.

Quick Intro on different compute options

Amazon EC2

Amazon Elastic Compute Cloud

This service provides on-demand and scalable compute capacity on the cloud.



- If companies needed to host their own application, in the past they would need to procure hardware - servers, storage etc.
- They would need to setup the data center, provide networking cooling - All of this was an added capital expense for the application.

But with Amazon EC2 , you can create and manage virtual servers on the cloud.

You only pay for how much you consume.

You can create and terminate machines at any point in time.



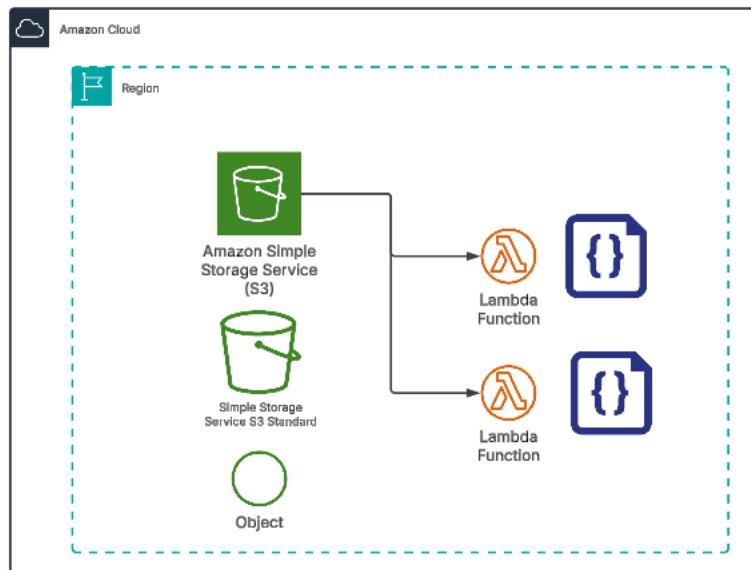
Within your AWS Account, you can choose a region.

Within the region , you have an Amazon VPC. The Amazon VPC is a logically isolated virtual network

The EC2 Instance is hosted within the Virtual Private Cloud.

AWS Lambda

This is a form of serverless compute



Here you don't need to manage the virtual servers. The infrastructure is managed for you.

You just host your code using AWS Lambda.

The Lambda function can respond based on an event. The Lambda service stops running the function once execution has completed.

You only pay for how much time the Lambda function ran.

Container-based applications

You can build container-based applications using tools such as Docker.



We can host the containers on Amazon EC2 Instances, but we have specialized tools and services that can be used to manage and orchestrate container-based workloads.



Amazon Elastic
Container Registry
(Amazon ECR)

This is a fully managed registry service that is used to host your container-based images.



Amazon Elastic
Container Service
(Amazon ECS)

This is a fully managed container orchestration service on AWS.



Amazon Elastic
Kubernetes Service

This is a managed Kubernetes service on the cloud. Here you can manage and orchestrate your container workloads using the Kubernetes software.

Generative AI

Introduction to Generative AI

Generative Artificial Intelligence

This is a subset of Artificial Intelligence and makes use of pre-trained models to generate text, images, videos etc.

Here the model has been trained on prior forms of data and understands how to generate new data.

So with Machine Learning, we train machine learning models for a specific purpose.

For Generating content, the model has been pretrained beforehand again, but this time, the purpose of the model is to generate content.

We all might be familiar with OpenAI's ChatGPT. This is a pre-trained transformer can be used to generate text based on our queries.

What can I help with?

What tools can be used on Amazon Web services for trying out the various foundation models.

+ Search Reason

↑

Create image Brainstorm Summarize text Analyze data More

AWS offers several tools and services for experimenting with foundation models and generative AI. Here are some key options:

1. Amazon Bedrock

- A fully managed service that allows you to build and scale generative AI applications using foundation models from multiple AI providers (like Anthropic, AI21 Labs, Stability AI, and Amazon's own Titan models).
- No need to manage infrastructure; you can integrate models via API.
- Provides customization options, including fine-tuning and embedding models in applications.

2. Amazon SageMaker JumpStart

- Offers pre-trained foundation models and allows you to fine-tune them on your own data.
- Includes built-in notebooks and training environments.
- Supports Hugging Face, TensorFlow, PyTorch, and other ML frameworks.

3. AWS Inferentia & Trainium

- Custom AWS chips designed for efficient inference and training of large-scale foundation models.
- Available through SageMaker and EC2 instances.

4. Amazon SageMaker Studio

- A complete IDE for building, training, and deploying ML models.
- Supports notebooks, automated model tuning, and data processing.

We can use the Dall-E model available with OpenAI to generate images.

A cat learning to be a cloud engineer with Amazon Web services

That sounds like an awesome idea! Imagine a fluffy cat sitting at a desk, staring at multiple monitors with AWS documentation open, a cloud architecture diagram on a whiteboard behind it, and a coffee cup that says "Purrfect DevOps." Maybe the cat is wearing little glasses and typing on a keyboard with a terminal open showing AWS CLI commands.

Do you want me to generate an image of this? Let me know if you have any specific details in mind!

↶ ↷ ↹ ↻ ↺

Yes



I can use GitHub Copilot to help generate code or troubleshoot existing code. It's like an assistant or buddy that can help me during my coding process.

Please generate a python script to use a foundation model with Amazon Web Services Bedrock

GitHub Copilot

To use a foundation model with Amazon Web Services (AWS) Bedrock, you need to follow these steps:

1. Set up AWS SDK for Python (Boto3):

```
pip install boto3
```
2. Write the Python script:

```
import boto3
import json

# Initialize a session using Amazon Bedrock
session = boto3.Session(
    aws_access_key_id='YOUR_AWS_ACCESS_KEY_ID',
    aws_secret_access_key='YOUR_AWS_SECRET_ACCESS_KEY',
    region_name='YOUR_AWS_REGION'
)
|
# Initialize the Bedrock client
bedrock_client = session.client('bedrock')

# Define the input for the foundation model
input_text = "Your input text here"

# Define the parameters for the foundation model
```

Ask Copilot

@ GPT 4o ▶

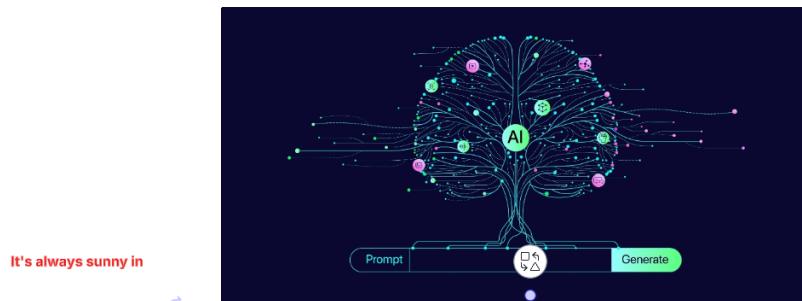
Large Language Models

Large Language Models

This is a type of machine learning model which is specifically designed for tasks which involve natural language processing.

Here the model is trained on large amounts of data and are normally used to generate text.

A large language model is used to predict the next word based on the set of text as the input.



The Large Language model would generate a set of words as the likely output with each output word having its own weightage.

Philadelphia

When it comes to building a large language model, we need to train it on a large data set.

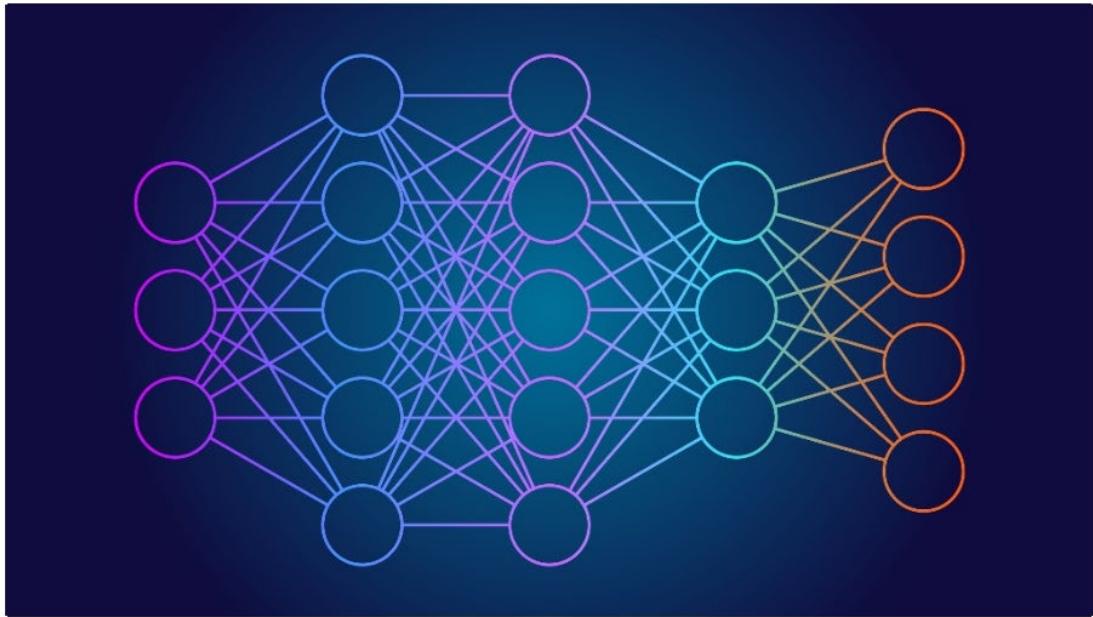
Imagine giving the ability of a model to be able to complete or generate text based on the input. It needs to be trained on all of the possible combinations on to complete sentences, complete text etc.

You need large amount of compute, you need large amounts of data , and you need time to build the large language model.

When it comes to compute , Graphical Processing Units (GPU) are used for training the model. GPU's allow for parallel processing and it makes it more efficient to train the models.

Then we have the GPT which is a generative pre-trained transformer. This is a type of large language model.

It is an artificial neural network that is based on deep learning. Here the learning is based on large data sets and unlabeled data.



ChatGPT is a generative artificial intelligence chatbot that was developed by a company known as OpenAI. Currently it is based on the GPT-4 large language model. The GPT , the transformer is what OpenAI developed to power ChatGPT.

Here the Chat part is just a friendly interface that allows us to interact with the Large Language model.

If you look at GPT-3 , the input data that used was around 45 TB. The data was filtered and then used to train the model - There is no actual stat for this.

GPT-4 was trained on around 13 trillion tokens.

If you were to train GPT on the data with just a single GPU it would take years and years to train the model.

Hence you would need thousands of GPU's running in tandem working to train the model.

What is a Foundation Model

Foundation Model!

These are large-scale machine learning models.

These models have already been trained on large data sets. These models can be used for tasks such as computer vision, natural language processing.

These foundation models are based on generative AI. Here it generates output based on input.

Different terms when it comes to foundation models

Text Generation - Here the model generates text based on an input text.

These models can be used to generate casual text, to generate text based on instructions. There are models that can also be used to generate code.

Image Generation - Here the model generates images based on text input.

Multi-Modal Model - Here the model can process different types of input - text, video, audio and image.

Embedding Model - Here the machine learning model uses data like text, the training data and converts it to a vector space. This helps the model to understand the semantics of the text and find similarities in the data set.

Word Embedding Models - Here the words are converted to vectors and similarities are drawn between the words - e.g - Word2vec.

BERT (Bidirectional Encoder Representations from Transformers) - This is another example of a word embedding model. The embeddings in the vector space are built based on the context of surrounding words.

Multi-Modal Embedding Models - Here the model can feed off from different types of data - Data in word documents, data within graphs etc.

Some of the Foundation Models

Anthropic Claude

Anthropic is an American-based AI safety and research company.

They built the foundation model of Claude that a competitor for OpenAI GPT model.

The model can be used for performing complex cognitive tasks, Vision Analysis, Code generation etc.

There are three versions of the model

1) Haiku - This is the fastest model available.

2) Sonnet - This provides a combination of performance and speed.

3) Opus - This can be used to perform highly complex tasks.

OpenAI

OpenAI has a set of models on offer.

1) GPT-4o model - This is a high-intelligence model. It takes in text and images and produces text as output.

2) o1 model - These are based on reinforcement learning and is used to perform complex reasoning.

3) GPT-4o audio - This accepts audio inputs and gives audio outputs.

4) DALL.E - This is used to generate images from text.

Stability.ai

This is a UK-based artificial intelligence company.

Stable Diffusion is a Gen AI model that can be used to generate images from text and image prompts.

- 1) Stable Diffusion - This is used to generate life-like images.
- 2) Stable Video Generation - This can be used to generate videos.
- 3) Stable Audio Generation - This can be used to generate audio.

Hugging face

This is a platform that hosts a tonne of models, data sets and demo applications. It's open source and publically available.

There is a specific collaboration in place between AWS and Hugging face.

You can use AWS services to deploy the various models available on Hugging face.

Meta

There are many models available from Meta.

Their most famous model is the Llama model.

This is meant to compete with the OpenAI GPT and Google Gemini model.

This is a large language model, you can use this to generate text based on input text.

Currently this is free to use.

Amazon

Amazon also has their own set of models.

Amazon Nova understanding models - These can be used to parse text, images, videos and generate text output.

Amazon Nova content generation models - These can be used to generates images and videos.

Amazon Titan - This is a large language model for text generation.

What is RAG - Retrieval Augmented Generation

Retrieval Augmented Generation

This is the technique wherein we use information from other data sources to supplement and improve the responses given by the model.



An LLM model is already trained on a large corpus of data. Here you can send prompts and queries and get a desired response.



But let's say that you want the model to supplement the responses with additional data based on your company's data stores.

Your company might be having its own internal data stores. You might have documents in Microsoft SharePoint. Obviously the LLM is not trained on your company data.

But now you might want the LLM to also consider the enterprise data you have when giving responses.

One option is to fine-tune the model wherein you train the model based on the enterprise data.

Or you use Retrieval Augmented Generation or RAG to augment the responses given by the model based on the data in your enterprise data stores.

RAG is very useful when you want to perform semantic searches on your data.



Amazon Simple
Storage Service
(S3)



Documents

Let's say that your company stores documents within the Amazon S3 service.

The documents format could be - .txt, .html, .pdf, .doc/.docx,.csv

And your LLM needs to make text-based searches within these documents.

For this you would make use of Embedding wherein your text is converted to
Vector Embeddings



This makes it easier to perform a semantic search over your enterprise unstructured data that is stored in the documents.

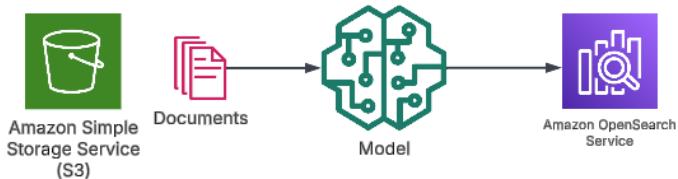
Lab - Amazon Bedrock - Knowledge Base - Implementation Overview



In order to make use of the Retrieval Augmented feature, we need to create a knowledge base in Amazon Bedrock.

We can create a knowledge base of unstructured data. The data can reside in Amazon S3, Microsoft SharePoint, Confluence etc. The supported data sources are provided in the AWS documentation.

In terms of unstructured data there is support for documents that contain text and images. You can have markdown , HTML or PDF documents as well.



We then use an embedding model which converts the data into vector embeddings in a target data store.

In terms of the Vector Embedding store there is support for - Amazon OpenSearch Serverless, Amazon Neptune, Amazon Aurora, PineCone, MongoDB Atlas, Redis Enterprise Cloud.

Once the ingestion process is complete, we then have our knowledge base in place.

Security and Monitoring on AWS

Identity and Access Management

Identity and Access Management

This is a managed service on AWS that allows you to control access to your resources.

You can create identities such as IAM users. You can then give permissions to the users over resources in the AWS account.

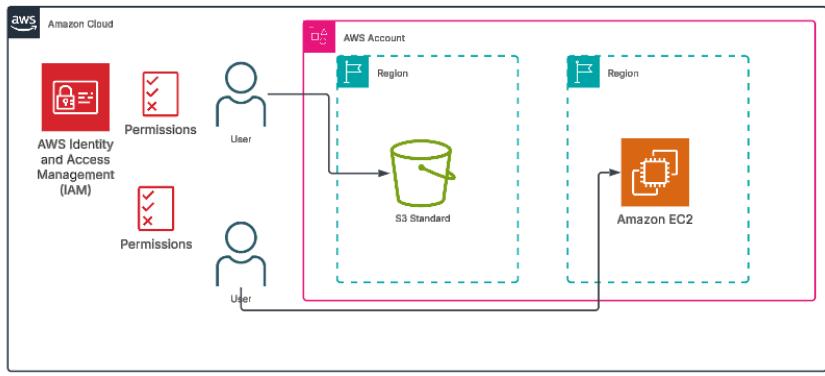
AWS IAM is responsible for both authentication (Here the user verifies that they are who they say they are) and authorization (Here the permissions to resources are verified)

Permissions can also be granted for one service to access another service.

Let's say an application hosted on an EC2 Instance needs to access the S3 service, you can assign permissions so that the EC2 instance has permissions to securely access the S3 resource.

AWS IAM as a service is available at a global level. It's available to all customers. It's a highly available and reliable service.

The service is free to use.

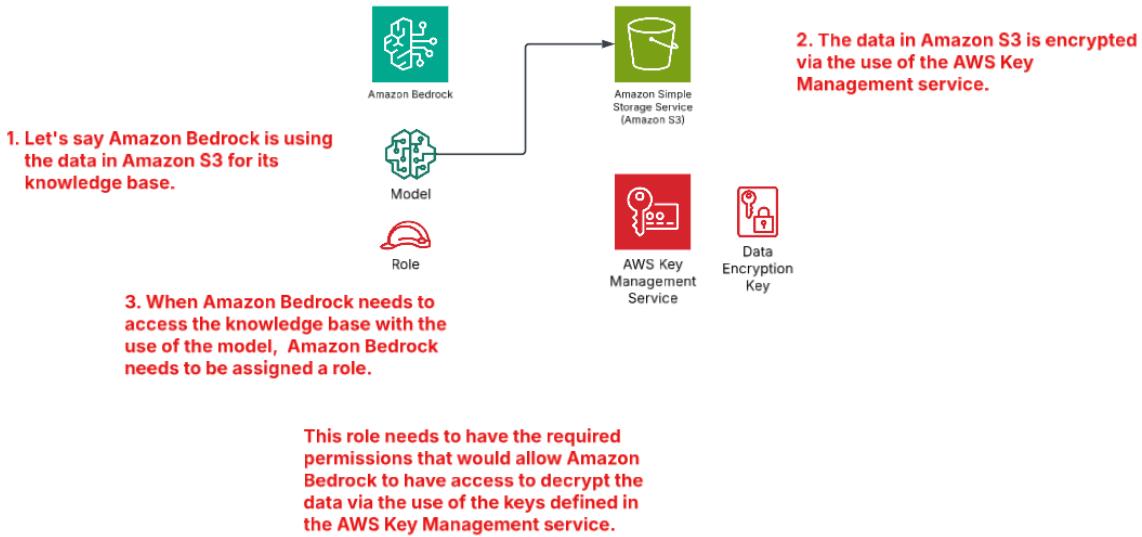
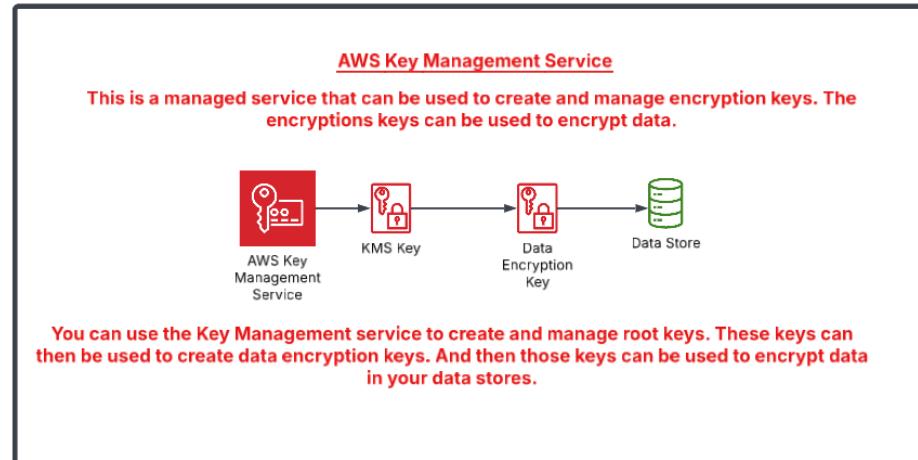


When a principal needs to access a resource in AWS, it sends a request to AWS , which gets passed onto AWS IAM. The request would contain the following details

- 1) Actions - The operations that the principal wants to perform - e.g Access an object in an S3 bucket.
- 2) Resources - What are the resources that the principal wants to access - e.g Objects in an S3 bucket.
- 3) Principal - The details of the principal who needs acces to the resource.
- 4) Environment data - Any sort of other information like the IP address of the device the user request is coming from.
- 5) Resource data - Any data specific to the resource.

Permissions are given via the use of policies. These policies contain the required permissions. These policies are nothing much JSON-based documents.

AWS Key Management service and Amazon Bedrock



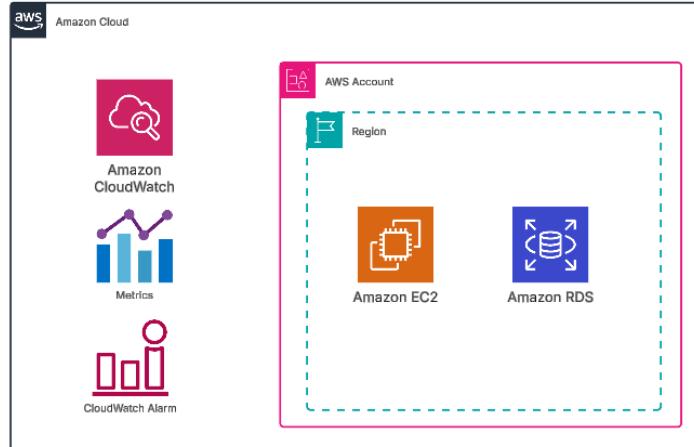
What is Amazon CloudWatch

CloudWatch - This service allows you to monitor your AWS resources and applications.

Here you can collect and track metrics for your AWS resources.

For example it's let's you track the CPU utilization, Network Utilization for your EC2-based resources.

You can also define alarms - If you want to be notified whenever a metric goes beyond a particular threshold you can make use of alarms to notify you accordingly.



CloudWatch also has network and internet monitoring capabilities.

Here you can monitor network traffic between your AWS resources, for example between your EC2 Instances. You can monitor for packet loss and other aspects when it comes to network monitoring.



We can also use a feature known as Amazon CloudWatch Logs

This allows you to collect logs across your AWS resources such as your EC2 Instances. You can have a central repository for your logs.

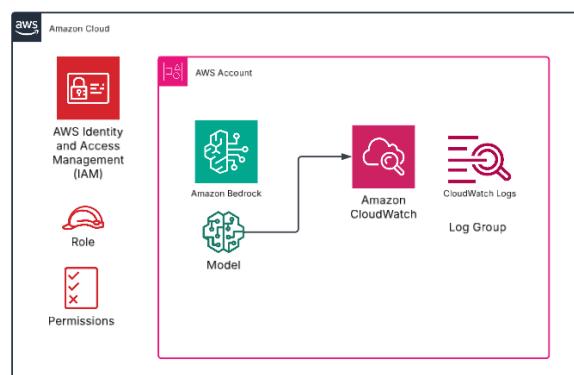
Amazon Bedrock and Amazon CloudWatch

2. In IAM we will first create a policy with permissions to write logs to the CloudWatch service.

3. Then we will create a role that can be assumed by the Bedrock service. We will attach the policy to this role.

4. Then from Amazon Bedrock, we can configure the monitoring feature when it comes to model invocation.

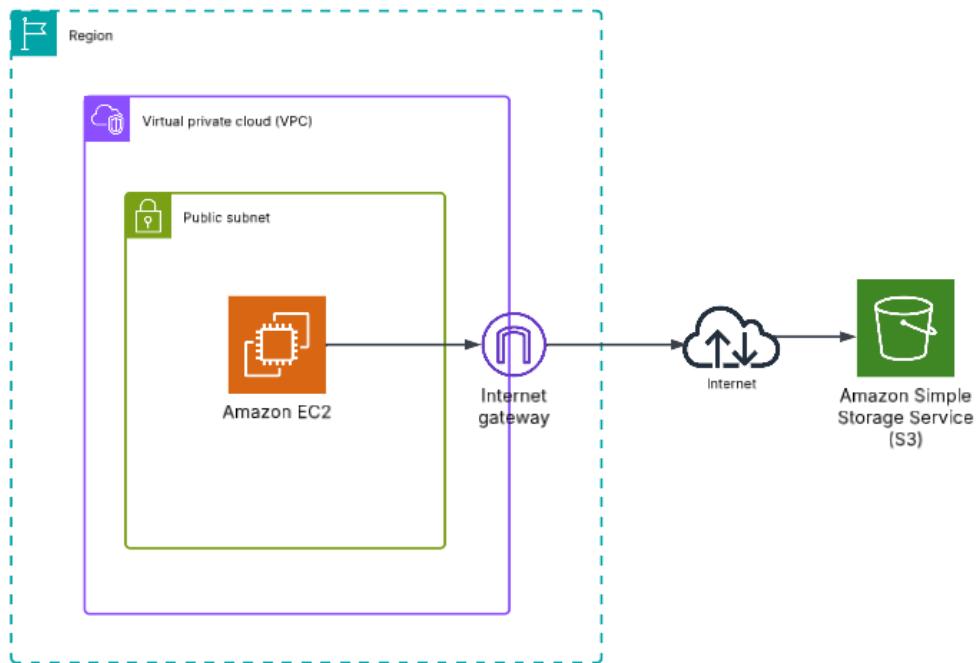
1. We need to create a log group for the logs which will be generated in Amazon Bedrock.



What we want to achieve - Whenever a model is invoked via Amazon Bedrock, we want to log each request and response.

We can do this via the model invocation monitoring feature in Amazon Bedrock.

Amazon Bedrock - AWS PrivateLink



Consider an application hosted on an EC2 Instance. Let's also consider that the application needs to access objects within an S3 bucket.

The EC2 Instance needs to be located within a VPC. The VPC is an isolated network on the cloud.

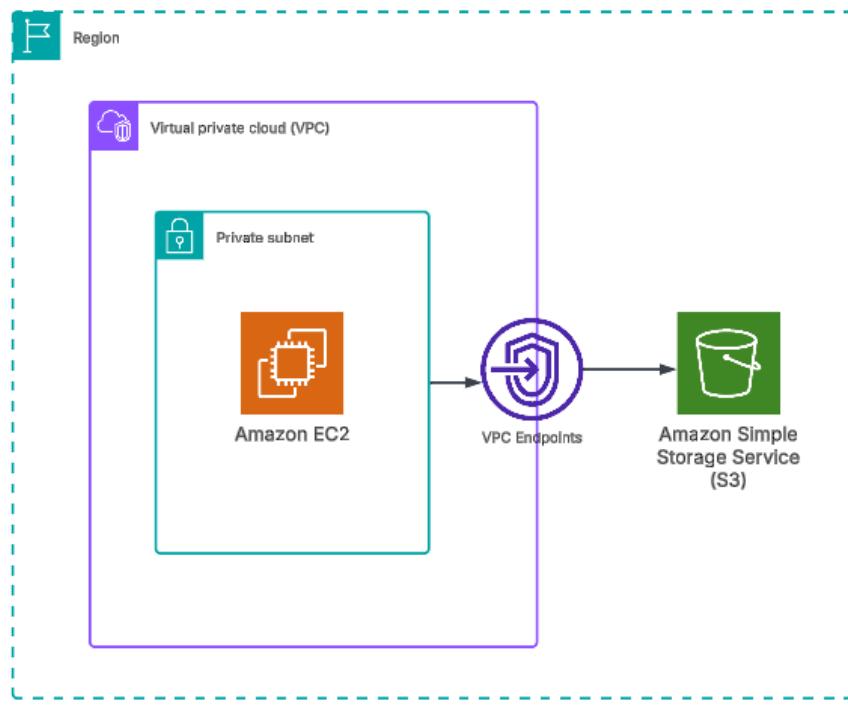
Amazon S3 is a public service. You access the service via the Internet.

In order for resources within a VPC to access resources on the Internet you need to have an Internet Gateway attached to the VPC.

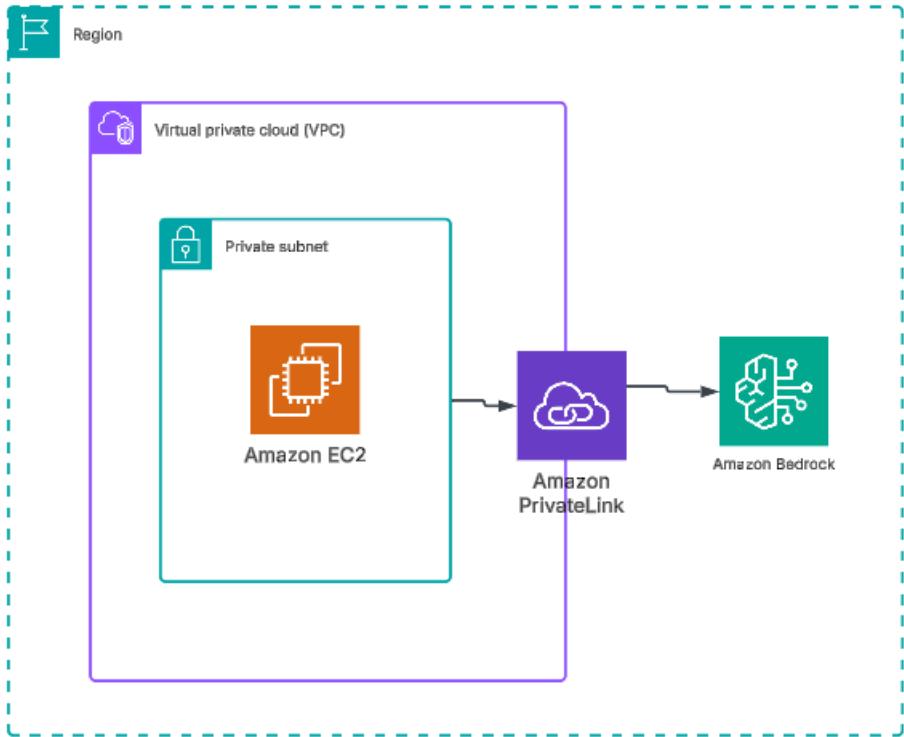
The Internet gateway allows communication between the Internet and the VPC.

VPC Endpoints

This feature allows resources within the VPC to securely and privately communicate with services on the Internet. Here the VPC resources don't need a Public IP address. Also the traffic between the Amazon VPC and the service does not leave the Amazon network.

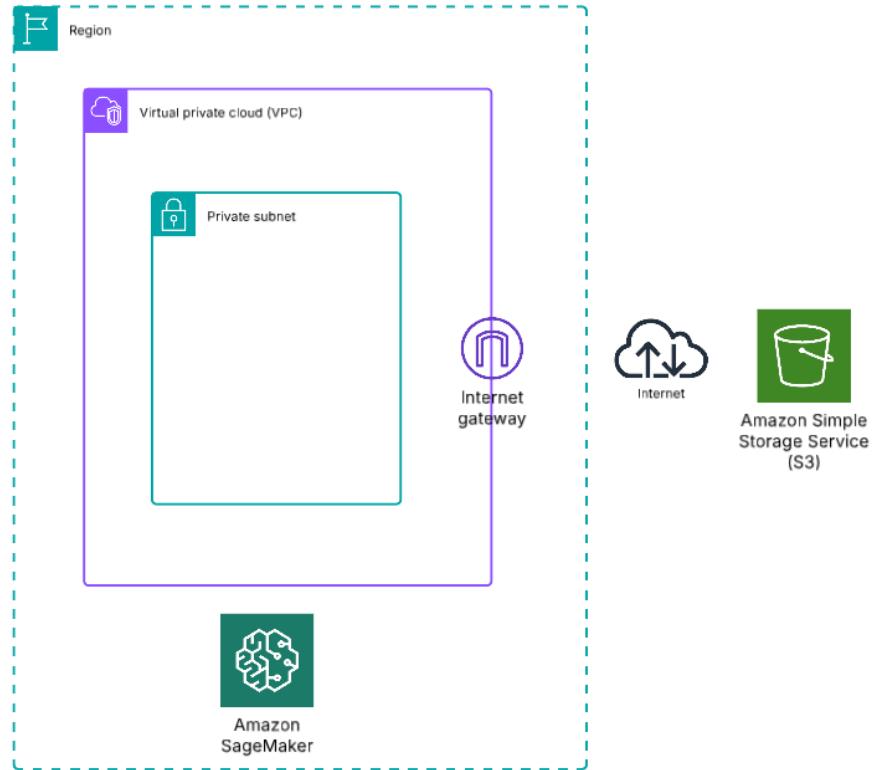


The VPC Endpoints are powered by AWS PrivateLink. The same AWS PrivateLink can be used to establish a secure and private communication between the VPC and Amazon Bedrock.



Amazon SageMaker and network isolation

When we work with Amazon SageMaker, the service will create a VPC for managing compute and storage resources. If there are data requests that need to reach Amazon S3 via the Internet, they pass through the Internet Gateway.



But for Amazon SageMaker Studio and the relevant notebooks, you can disable Internet access by mentioning a VPC only network access when onboarding Studio.

Also when you make use of training and deployed inference containers, they are enabled by default to work with the Internet.

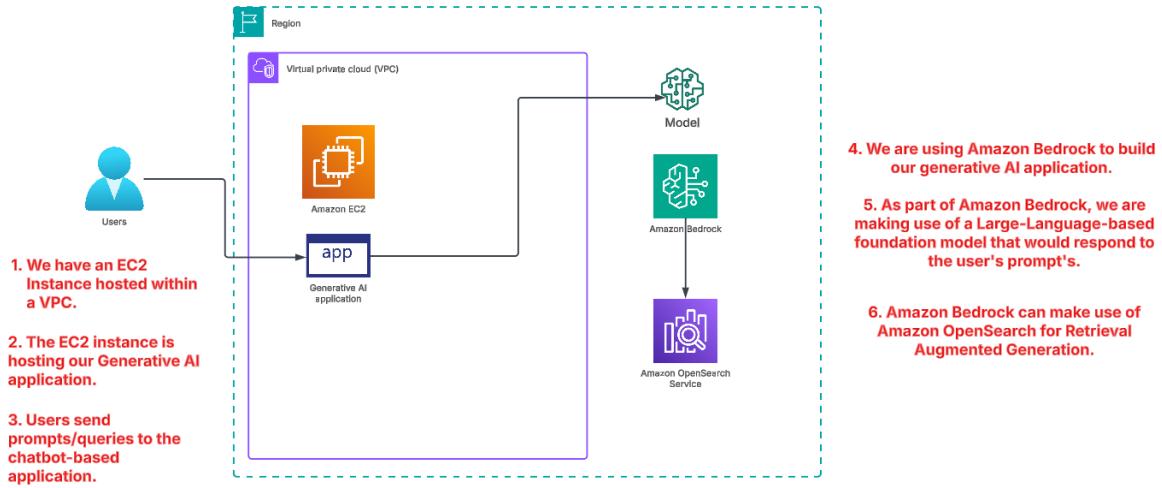
But you can also enable network isolation for the containers. Here the containers cannot make any outbound network calls even to Amazon S3.

Quick note on the design of a conversational chatbot

You want to build a conversational assistant , a chatbot like ChatGPT. Even on a lot of websites, you have chatbots that provide customer support.

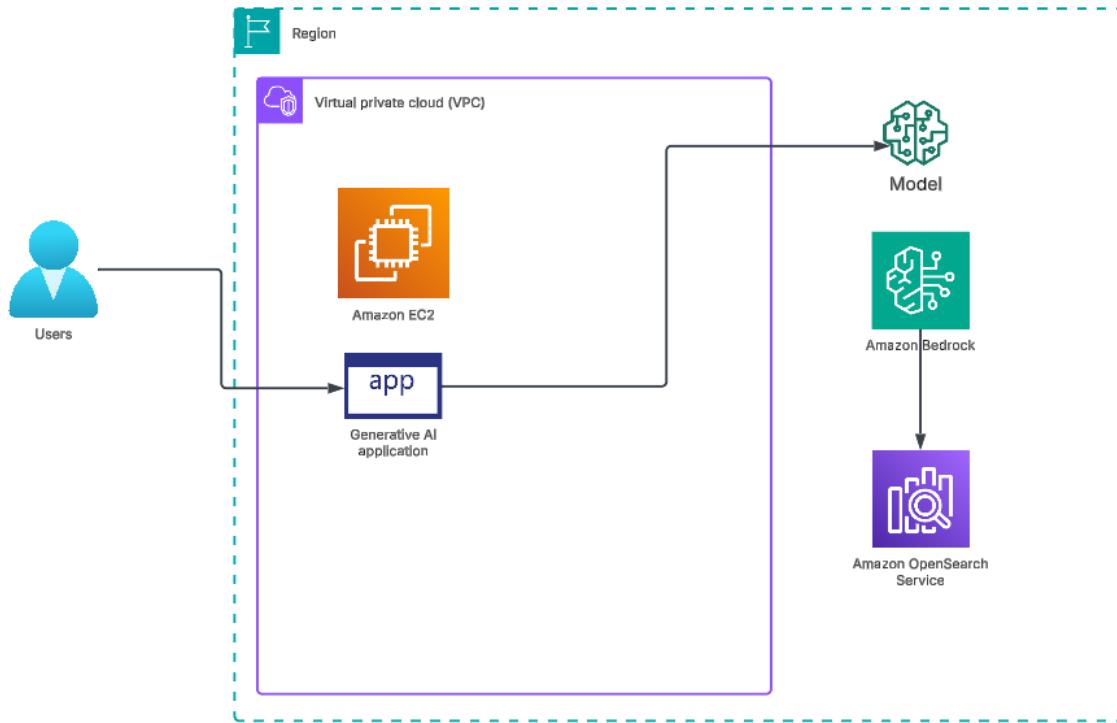
You could create a chatbot that can take in text and provide answers in text. A chatbot could be programmed to take in images.

Based on whatever we have learnt , we can implement the chatbot as follows



Securing your Gen-AI applications

Securing your Gen-AI applications



We need to have a secure environment when it comes to our Gen-AI application.

Malicious users can send harmful prompts to try to get undesired or sensitive information.

We can use Amazon Bedrock Guardrails to apply topic and content filters and lot more to protect the prompt data that gets sent to the large language model.

But then we need to be aware of the different threats when it comes to creating prompts.

Prompt Injection - Here a malicious user can provide instructions to the original user prompt to make the large language model follow the malicious prompt instead of the original user prompt.

Prompt Leaking - Here a malicious user make a model reveal its internal instructions on how it is supposed to respond to users.

Jailbreaking - Here a malicious user can send prompts in such a way that it makes the model produce unintended outputs.

Prompted persona switches - Here a malicious user give instructions to the model to change its context in the way it responds. This may cause the model to give unintended outputs.

Extracting the prompt template - Here a malicious user asks the model to provide its initial instructions. This would give more information to the malicious user on how the model was configured to answer user queries.

Using Amazon Guardrail can help protect against the above mentioned attacks.

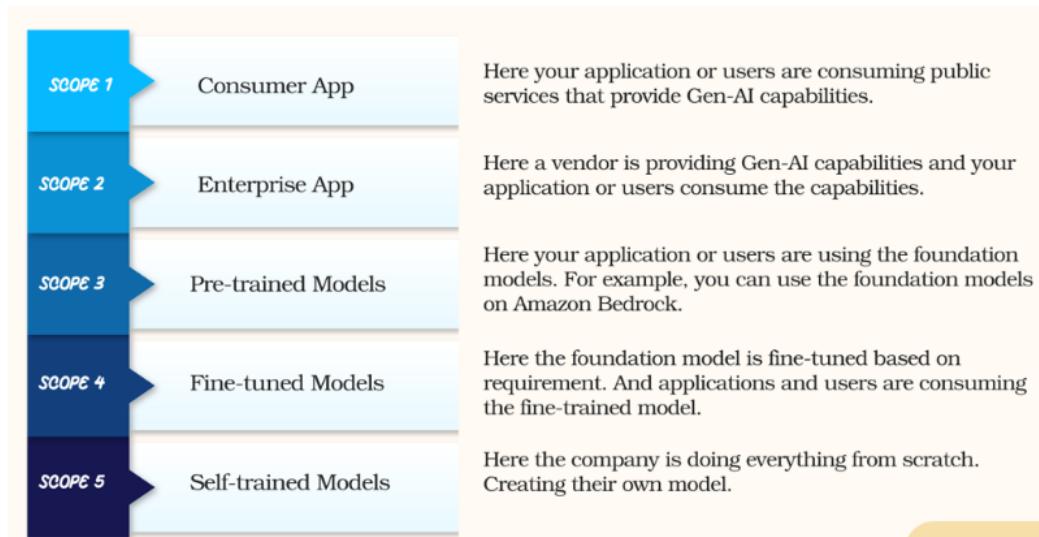
Train the LLM to detect specific prompt attacks

Generative AI Security Scoping Matrix

Generative AI Security Scoping Matrix

AWS has developed a security matrix when it comes to considering how to safeguard your Gen-AI based applications.

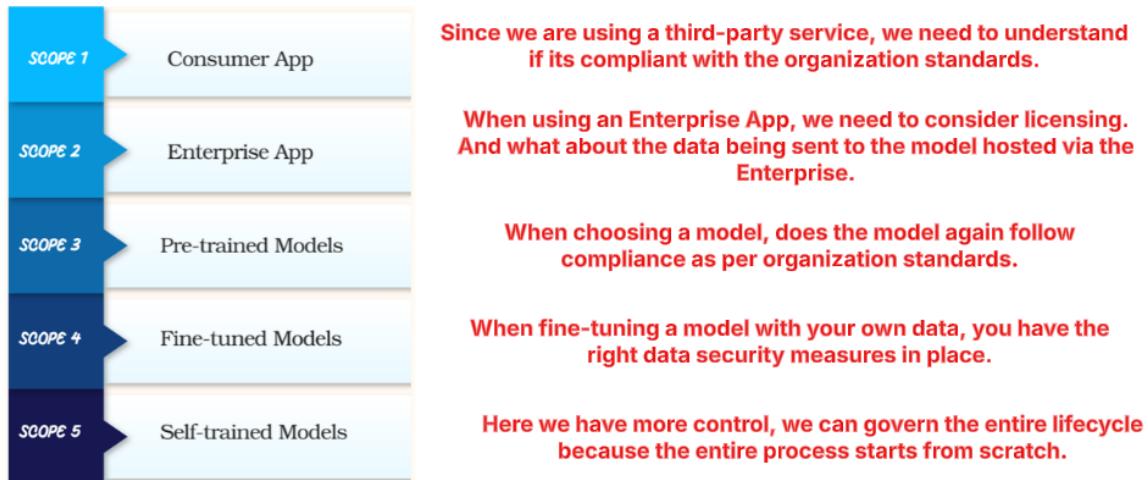
This does depend on several factors, first is to define the scope on how the Gen-AI application is being used.



We then have different security disciplines

We then have different security disciplines

Governance and compliance



Legal and Privacy

When looking into Scope 1 and 2, do we have the correct End-User agreements in place.

Regulations when it comes to usage of data for the Enterprise application.

Protecting privacy concerns of data when being used to find-train a model.

Controls

When using pre-trained and fine-tuning models, when using Amazon Bedrock, you do have various controls in place to secure the model and the data.

You can use AWS IAM to control access.

If you are using RAG to supplement company data for Scope 4, then how are using controls to protect the vector embeddings data store.

Risk management

For scope 1 and 2 , understand the risks with using a Public or Enterprise service for Gen-AI needs.

The Public vendor or Enterprise vendor has trained their model, you don't have control.

Manage how malicious users can use prompt engineering techniques to understand the model behaviour.

Resilience

What happens if the Gen-AI service provided by the public vendor or Enterprise App goes down.

When using Amazon Bedrock or your own complete environment as per scope 5, you have more controls to implement resiliency for your entire environment.