INTRODUCTION

AMAZON S3 - SIMPLE STORAGE SERVICE





AMAZON S3

This is an object storage service on AWS. Here the underlying storage is managed for you. The service provides scalability, data availability, security and performance.

AMAZON S3

General Purpose bucket

- In order to start uploading objects to S3, we must first create an Amazon S3 bucket. The bucket is used to hold the objects.
- You can store any number of objects within an S3 bucket.
- The name of the bucket needs to be unique. The name cannot be used by another AWS Account.
- You create a bucket in the region of your choice.
- In order to delete an S3 bucket, we first need to empty the bucket.

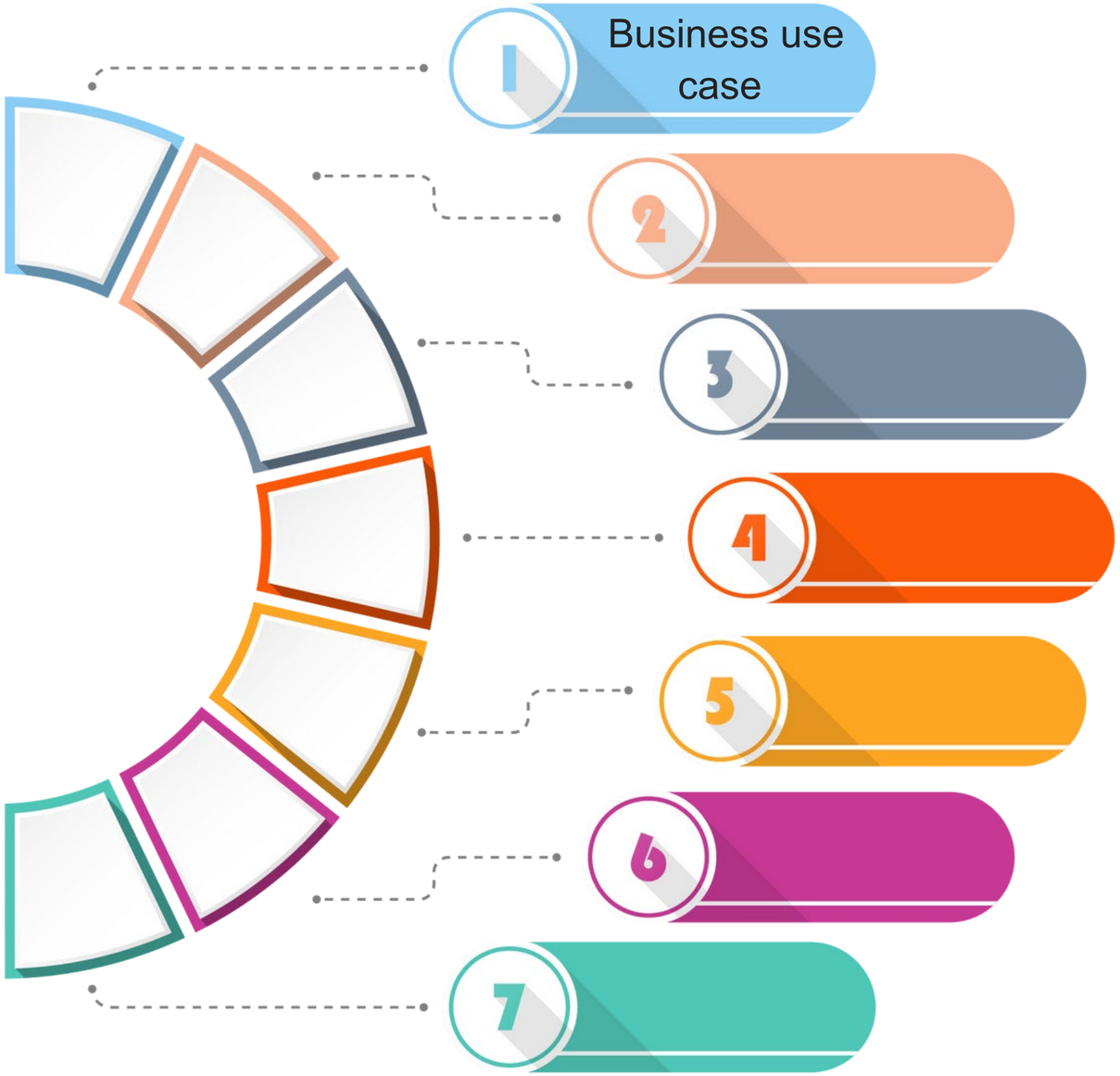
AMAZON S3

Objects

- Your data are stored as objects in an S3 bucket.
- Each object has a key - This is the name assigned to the object.
- Each object has a value - This is the content of the object.
- Each object gets a unique URL that allows you to access the object. But you need to have the right permissions to be able to access the object.

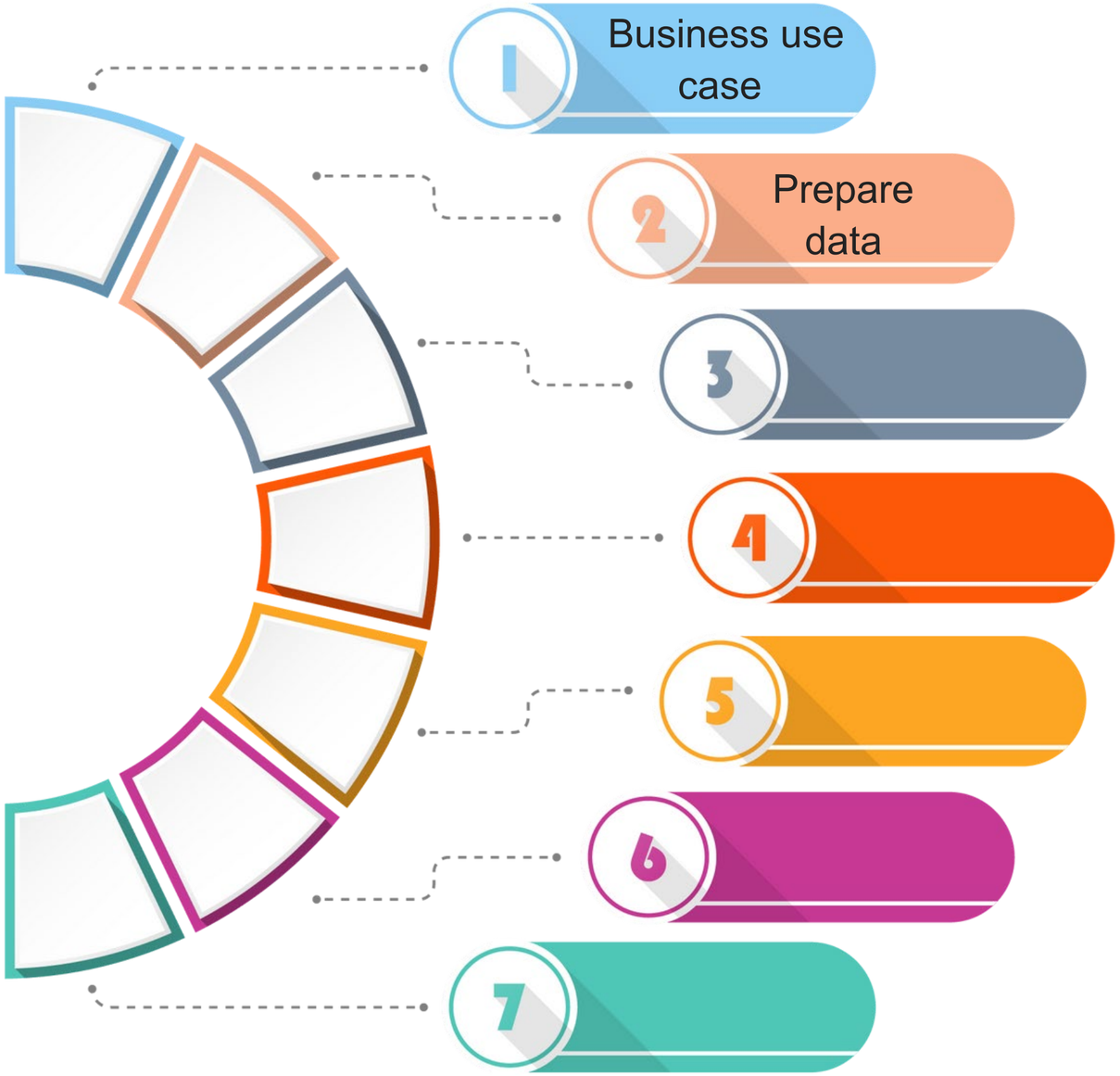
LET’S WORK ON MACHINE LEARNING

THE MACHINE LEARNING PROCESS



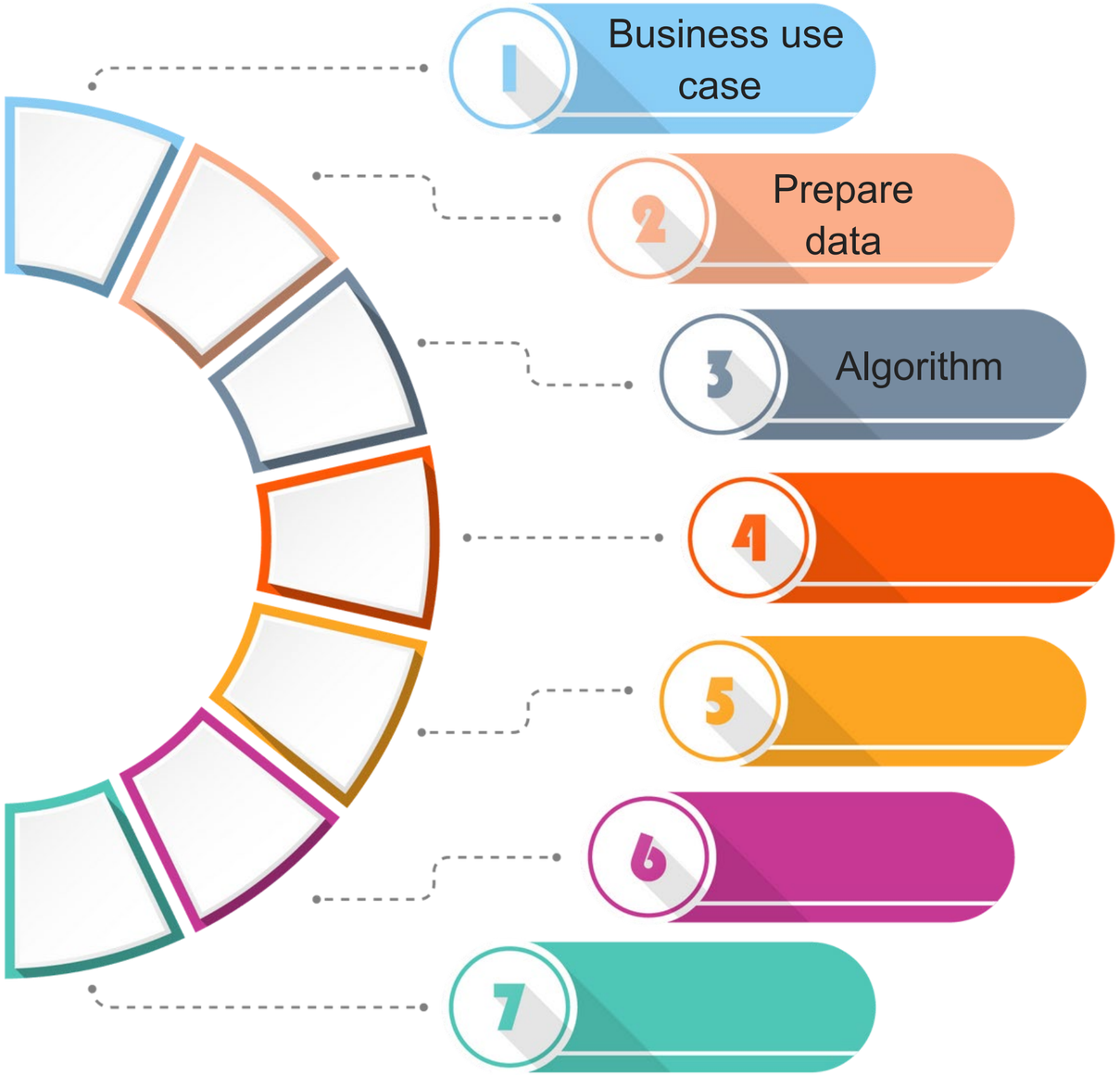
Business use case

- Why do we need to build the Machine Learning model?
- What business problem or requirement is it trying to solve.
- Have we justified the value behind building the Machine Learning model.



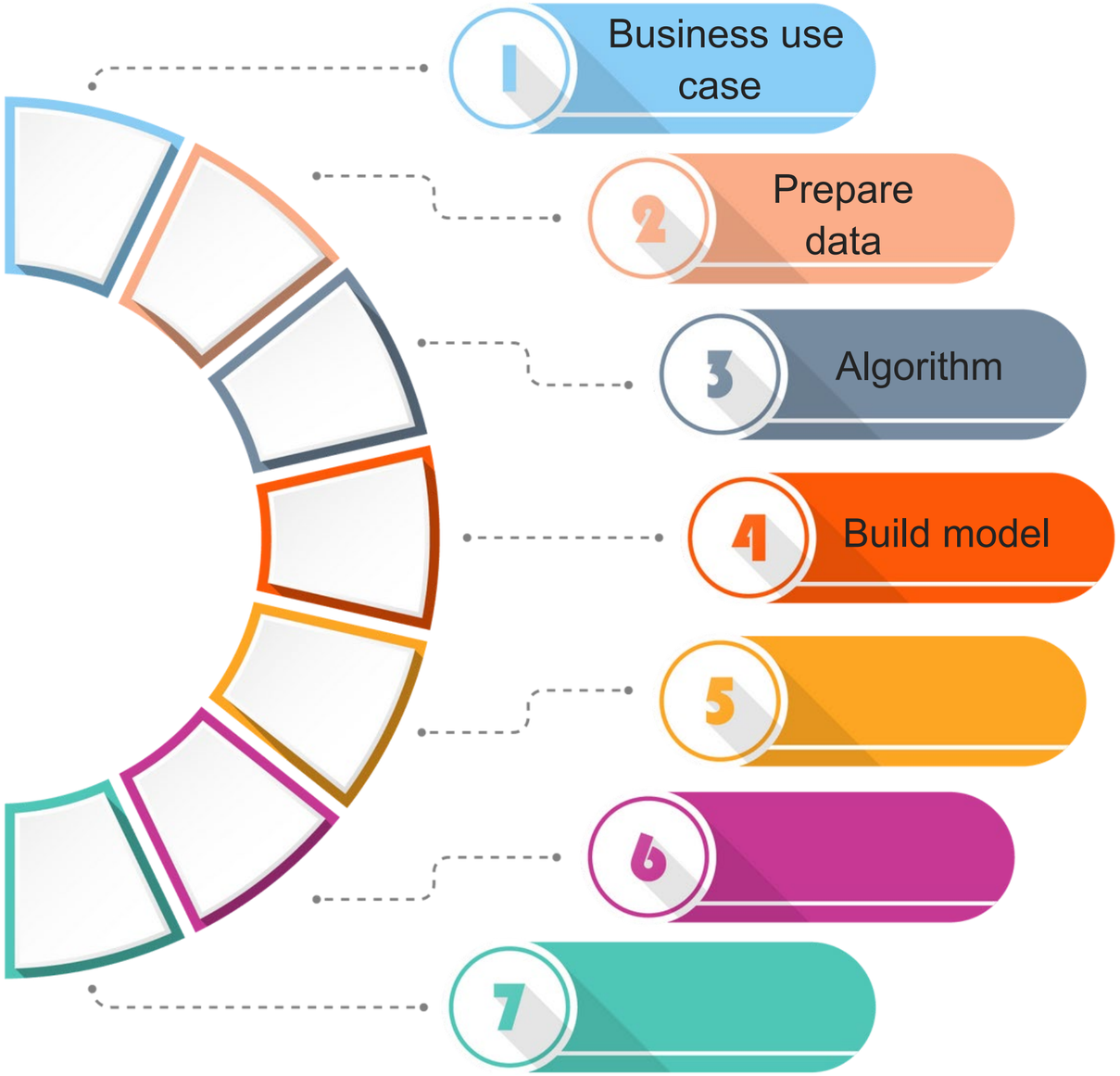
Prepare Data

- This is the most important step. Data is crucial to building a good Machine Learning model.
- We need to identify the sources of data. We need to cleanse and prepare the data.
- We need to extract a rich set of features from the data set and label the data if required.



Algorithm

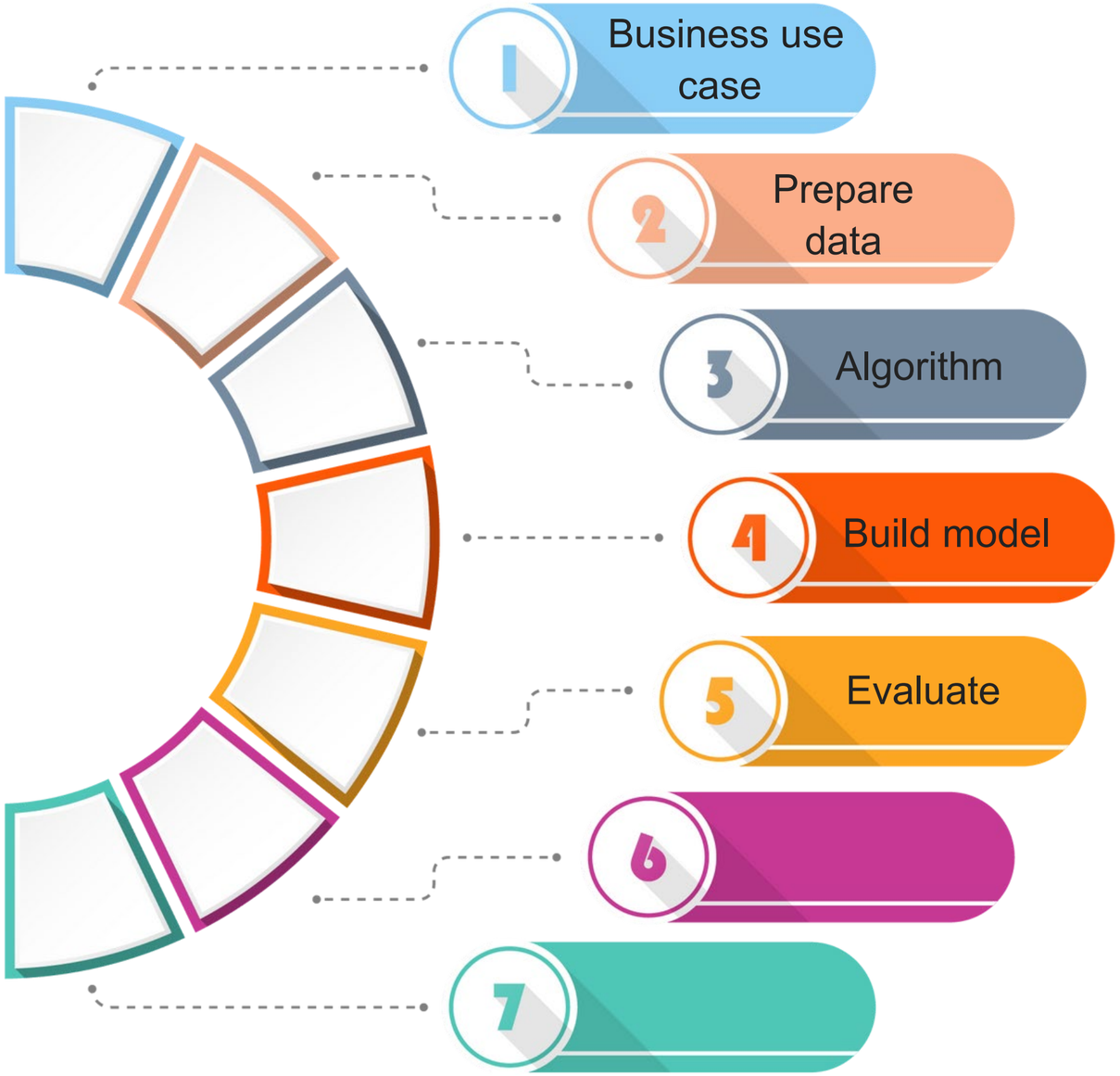
- There are many mathematical algorithms that can be used to build the Machine Learning model.
- It depends on the requirement of the model. Are we trying to forecast numerical values?
- Are we trying to classify data values?



Build Model

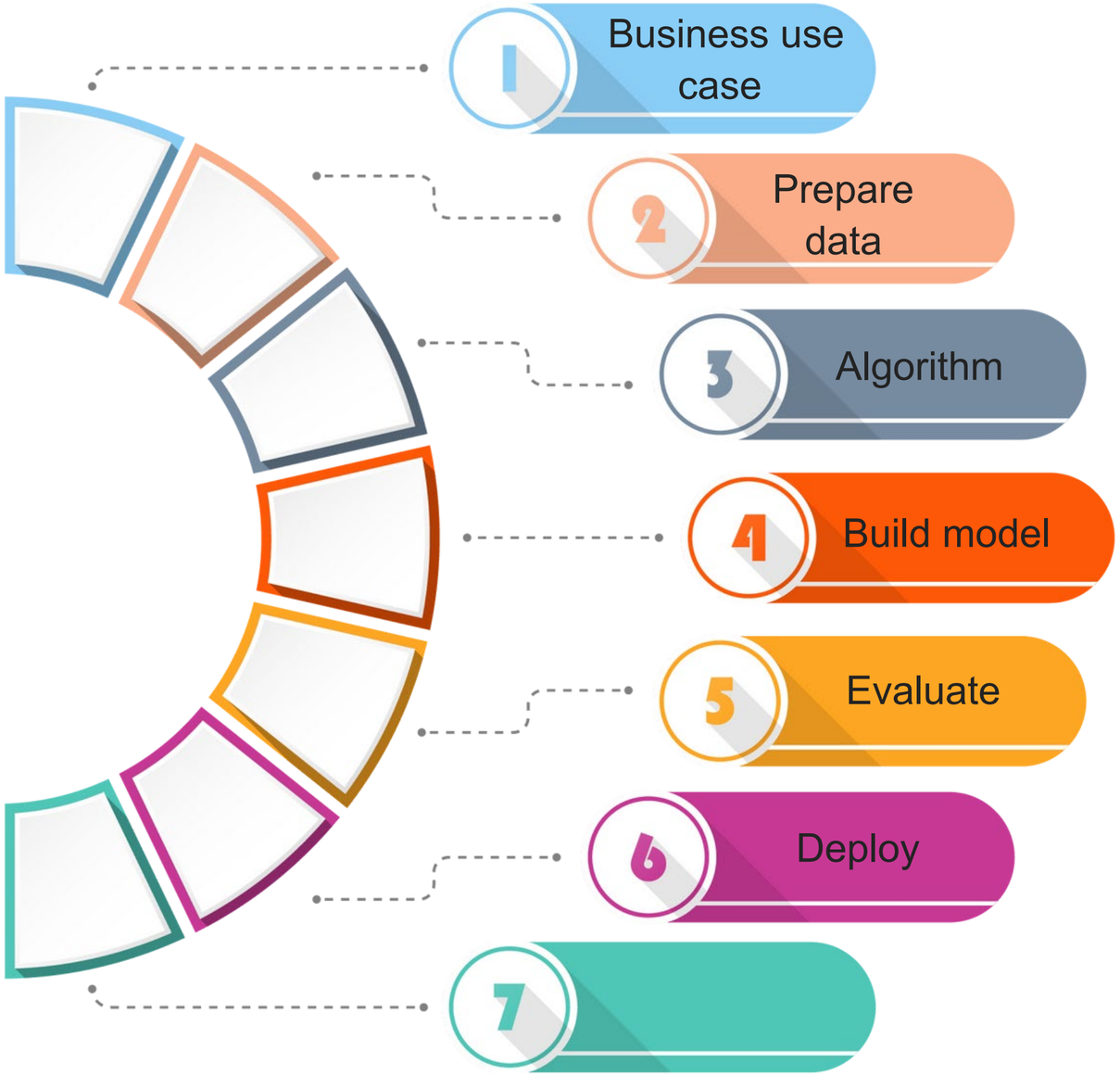
- Here we build the Machine Learning model.

- Using the Machine Learning algorithm and the feature -based data set, we set to build our Machine Learning model.



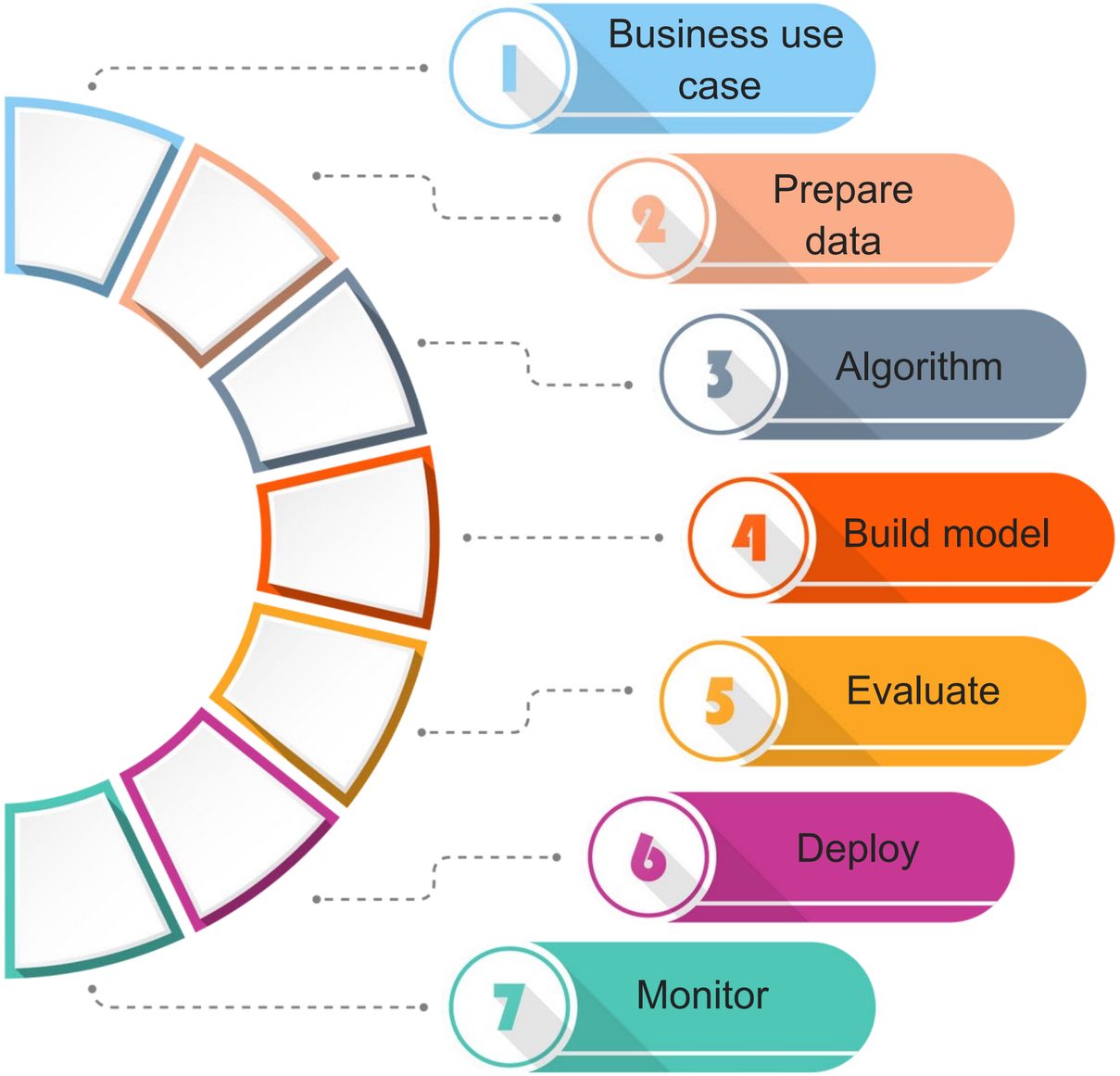
Evaluation

- After building our model, we need to evaluate the Machine Learning model.
- We can evaluate the model based on test data.
- Does the model accurately predict or classify the results correctly?
- Does the model have any bias?
- If there is a problem in the evaluation, we then we need to revisit our earlier steps - Maybe we need a larger data set, we need to choose maybe another algorithm.



Deploy Model

- We then deploy the Machine Learning model.
- The model can then be consumed by applications or users.
- We can submit input data and get the results accordingly.
- Does the model have any bias?



Monitor Model

- We need to continually monitor the model.
- In production is it giving the right results.
- Do we need to fine tune the model with more data. Maybe we need to train the model again.
- Maybe the input values have changed, a new feature needs to be added to the data set and we need to train the model again.

Roles

- Data Engineer - Here the engineer would be responsible for activities such as data cleansing, making sure the data is of use.
- Data Scientist - Here the responsibility would be to get insights from data, build the Machine Learning models.
- Machine Learning Engineer - Here the engineer would be responsible to deploy the Machine Learning models, look into the infrastructure for the model.

AMAZON SAGEMAKER AI

AMAZON SAGEMAKER AI

This is a fully managed machine learning service. Here data scientists and developers can easily build, train and deploy Machine Learning models.

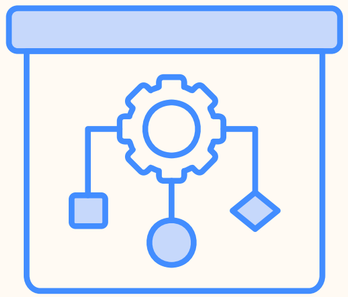
AMAZON SAGEMAKER AI

- What's our goal - Build a Machine Learning model.
- What do we need - Machine Learning Algorithm and a feature -based data set.
- We need tools and infrastructure to build and deploy the model.

AMAZON SAGEMAKER AI

- We need to use a tool to gather the initial data set.
- We need to use methods to clean and extract the required features from the data set.
- We need to also visualize our data from different aspects to understand - Do we have the right feature set, are there any missing values etc.

AMAZON SAGEMAKER AI



- We can use many pre-built machine learning algorithms. They are available from the Internet. They are also made available very easily in the tool.
- Then we need compute power. This power is required by the algorithm to go through the entire large data set to search for patterns.
- Trying to create a model with our laptop with a large data set is a challenge.

AMAZON SAGEMAKER AI

- Once the model has been built, we need to host the model. This is so that applications and users can call the model to make predictions.
- We again need compute power to host the model. We need an endpoint that users and applications can call.

AMAZON SAGEMAKER AI

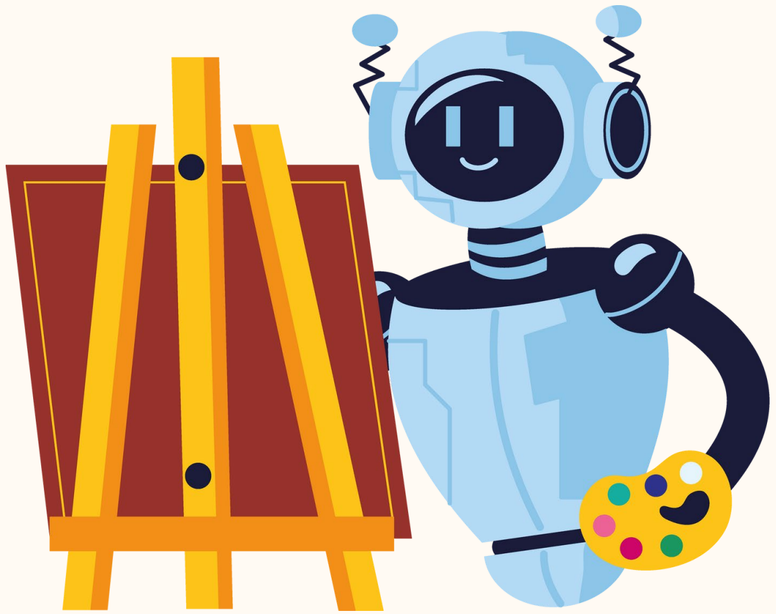
- To get started with Amazon SageMaker we first have to create a domain.
- We can create an Amazon SageMaker Unified Domain. This is an entity that is used to connect the various assests in SageMaker, the users and the projects.

AMAZON SAGEMAKER CANVAS

AMAZON SAGEMAKER CANVAS

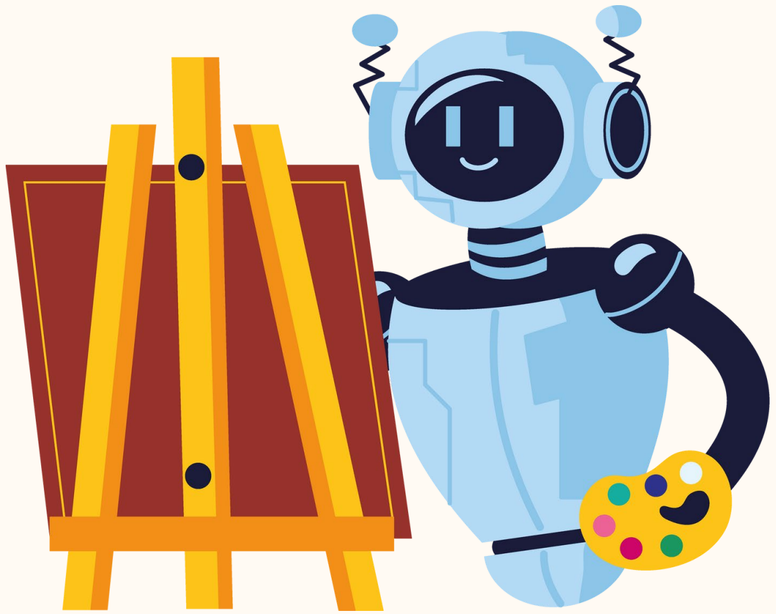
This service is a low-code option that allows you to use Machine Learning to make predictions. You can also use the tool to work with popular Large Language models.

AMAZON SAGEMAKER CANVAS



- This service is ideal for users who don't have any development experience.
- It allows them to import data from many data sources, build Machine Learning models, evaluate the model's performance and generate predictions.
- There is support for building regression models, binary and multi -class models, time series forecasting.

AMAZON SAGEMAKER CANVAS



- You can also chat with popular large language models.
- You can use pre-built models to get help with tasks such as generating content, summarizing and categorizing documents.

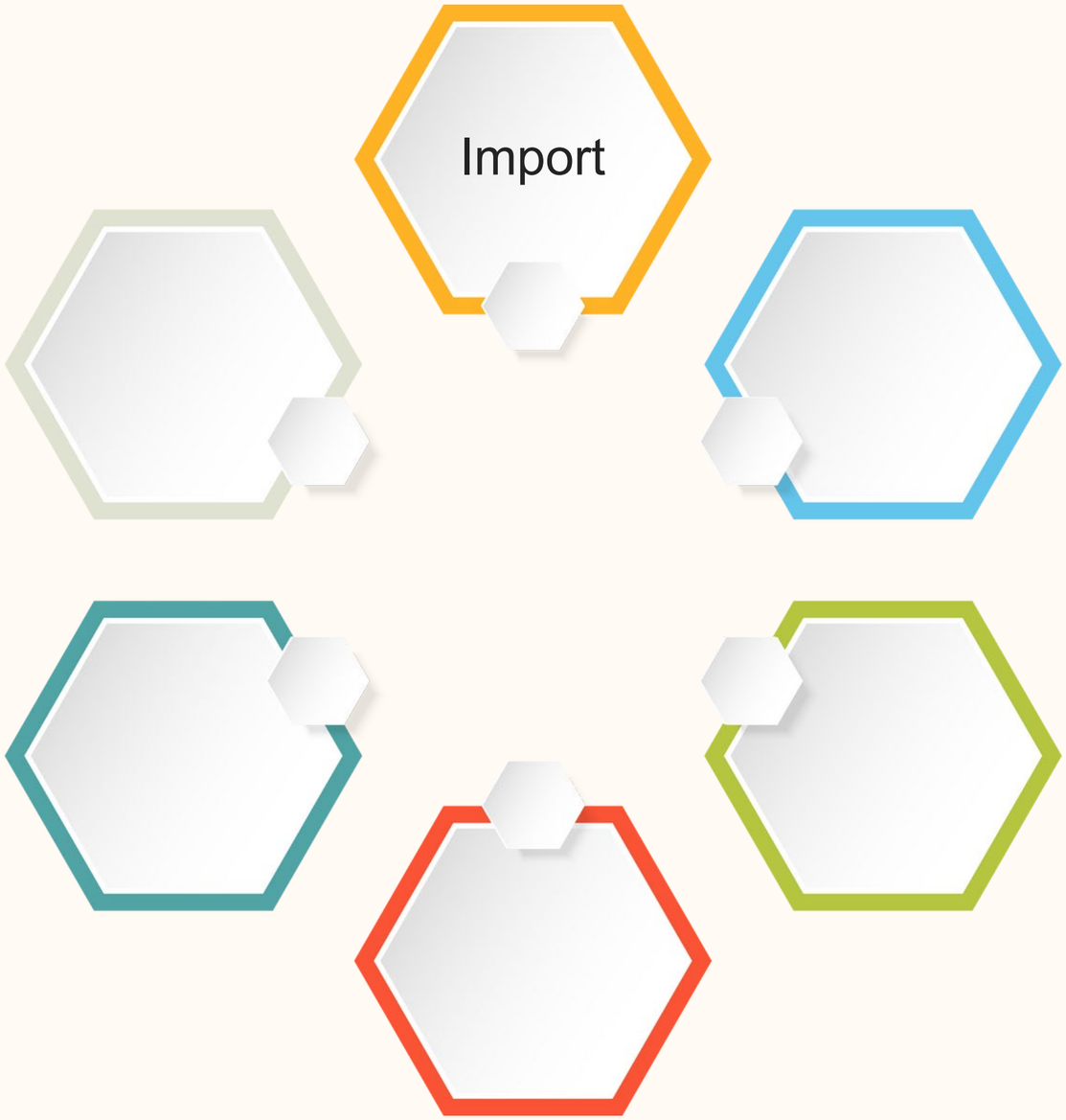
AMAZON SAGEMAKER DATA WRANGLER

AMAZON SAGEMAKER DATA WRANGLER

This service can be used to import, prepare and transform data. This data can then be used in your machine learning workflows.

AMAZON SAGEMAKER DATA WRANGLER

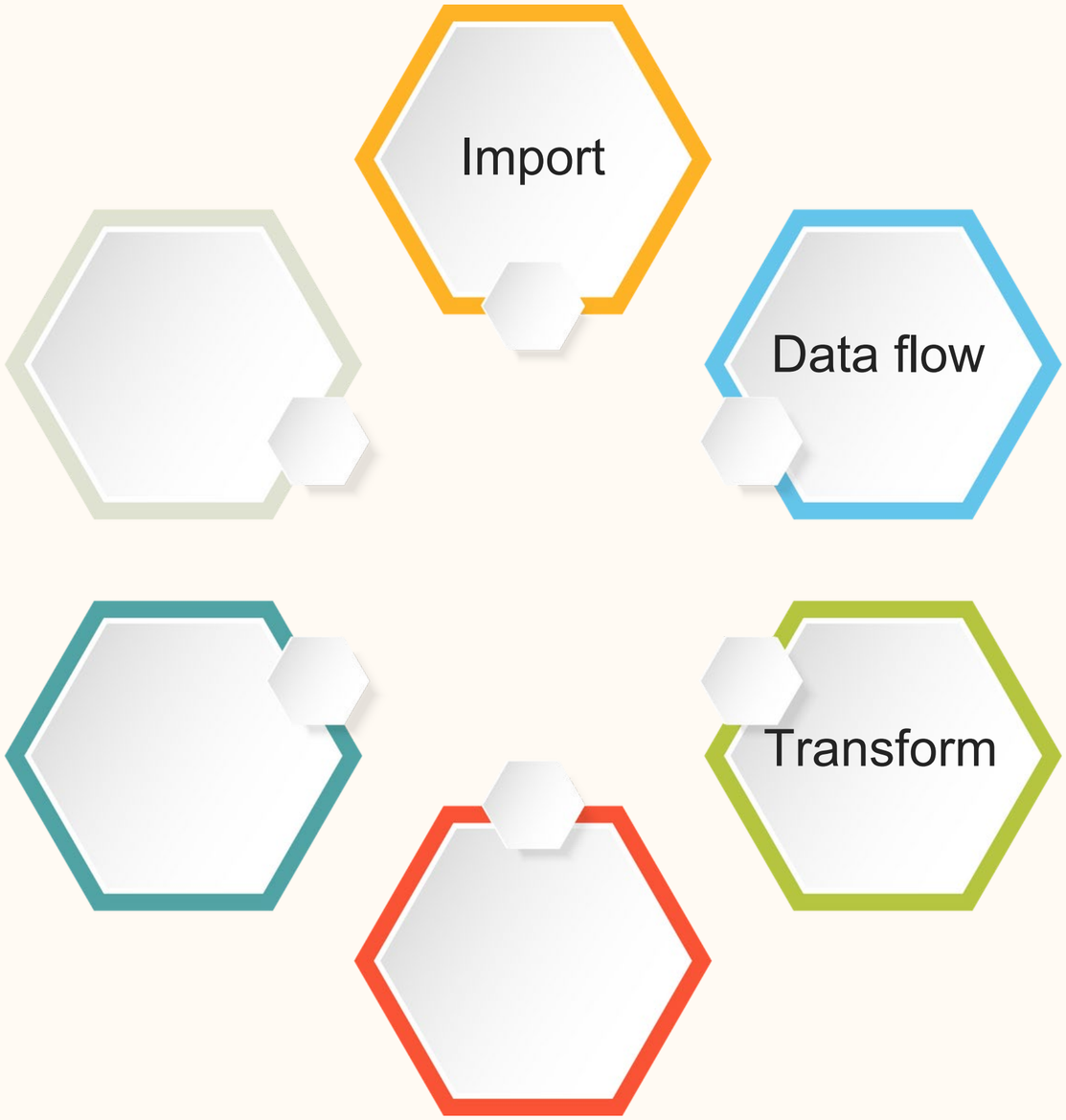
We can connect to different data sources and import data - Amazon S3, Amazon Athena, Amazon Redshift, Snowflake and Databricks.



AMAZON SAGEMAKER DATA WRANGLER

We can create a data flow that can combine data from different data sources.

AMAZON SAGEMAKER DATA WRANGLER



We can clean and transform data as required. We can cleanse data and prepare the data accordingly.

AMAZON SAGEMAKER DATA WRANGLER

We can generate data insights and get quality reports.

AMAZON SAGEMAKER DATA WRANGLER

We can analyze the data using different visualizations.

AMAZON SAGEMAKER DATA WRANGLER

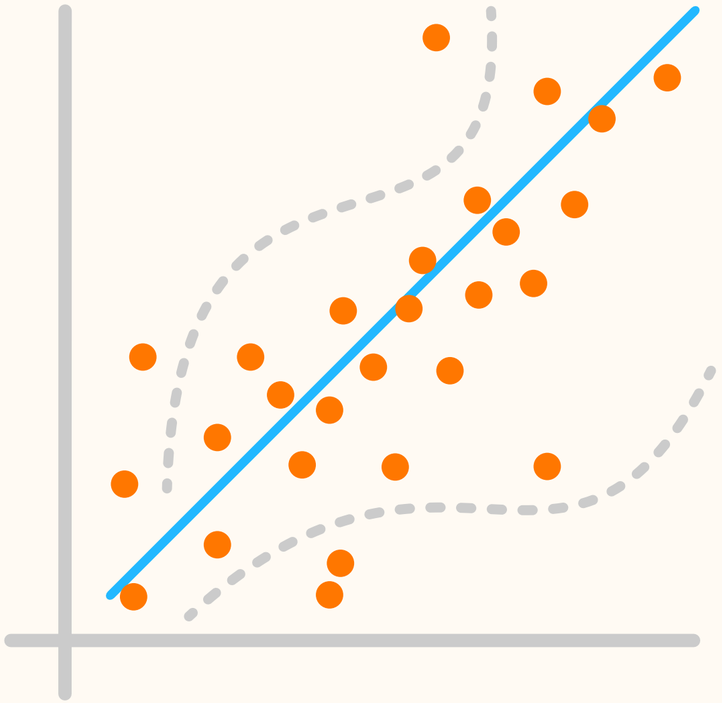
We can also export our data to Amazon S3, SageMaker pipelines, SageMaker Feature store.

MACHINE LEARNING ALGORITHMS

MACHINE LEARNING ALGORITHMS

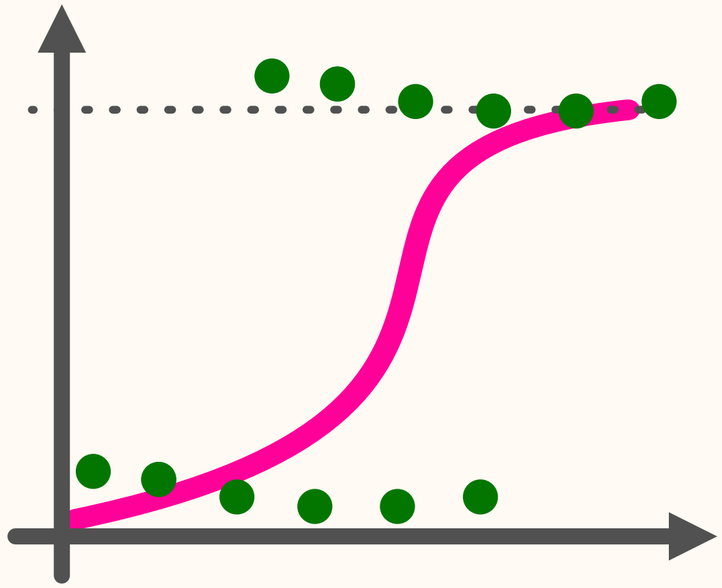
There are different Machine Learning algorithms. You can choose the required algorithm based on the problem you are trying to solve.

LINEAR REGRESSION



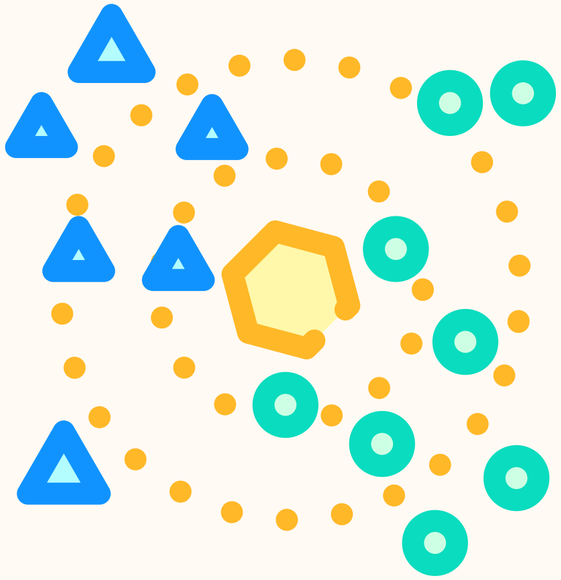
- Here the algorithm would try to understand the relationship between a scalar response and one or more independent variables.
- Linear equation : y = 2x + 5. Here we can have different values of y based on the input variable of x.
- The algorithm based on the input data would try to best find the linear equation that best fits the input and output values.
- This is more of a supervised learning algorithm.

LOGISTIC REGRESSION



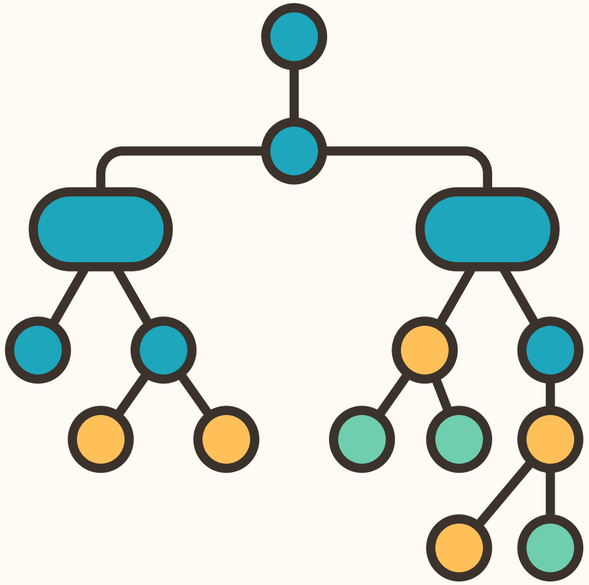
- Here the algorithm would try to determine the probability of an output event based on the input.
- It can use a logistic function to determine the output value.
- For example , let's say that you are using a Machine Learning model to determine if an email is spam or not. Instead of just a 0 or 1 to determine this, you want to determine based on the input the probability as to whether an email is spam or not.

K-NEAREST NEIGHBOR



- This algorithm is used for classification. Here it classifies a value based on its similar attributes to other values. This again falls into supervised learning.
- Here you have a distance metric between the points.
- The you define the value of k. This is the number of neighbors considered before assigning a value to a class. For example, if you assign the value of k as 1, then an input value will be assigned to the same class as its single nearest neighbor.

DECISION TREE



- A decision tree at its simplest form makes use of nodes that help to determine the output based on simple yes and no outcomes.
- For example, if you want to use a decision tree to determine the risk of having a heart attack, you can have nodes for different input values such as height, weight, age, does the patient smoke, other attributes.
- With a decision tree it also becomes easy to see all of the decision being made to arrive at the final result.

SUPERVISED LEARNING

Algorihtm	Description	Data Input type
Linear Regression	This is used to model the relationship between a dependent variable and one or more independent variables.	Numerical
logistic Regression	This is used to predict the possibility of an outcome - Normally used for binary classification.	Numerical/TTXT
Decision trees	Here you have different rules like the branches of a tree. Based on an outcome a particular branch is is chosen.	Numerical/TTXT
K-nearest neighbors	This is used for classification or regression and is based on the distance between points.	Numerical/TTXT
DeepAr forecasting	This is used for forecasting scalar time series values using recurrent neural networks	Time-series

UNSUPERVISED LEARNING

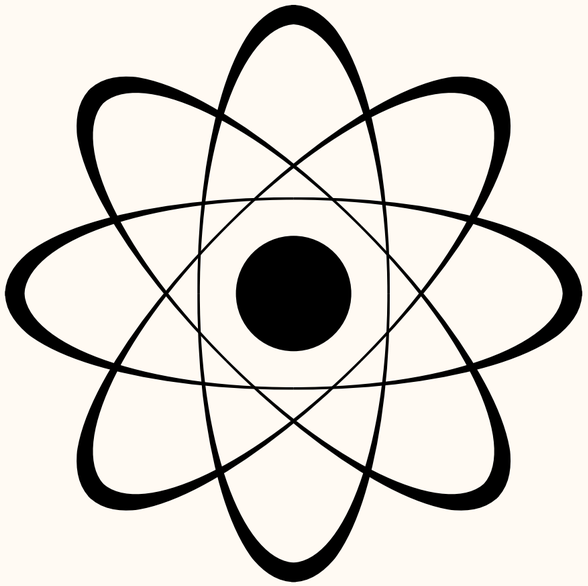
Algorihtm	Description	Data Input type
IP Insights	This is used to detect patterns in the usage of IP address.. You can see if a user is logging in from an anonymous IP address.	Entity - IP address
K-means	This is used for grouping of data. Here you can group data points that are similar to each other.	Numerical/TXT
Principal component analysis	This is used to reduce the dimensionality or the number of features within the data set. But at the same time, it maintains as much as information is required for training.	Numerical/TXT
Random cut forest	This is used to detect anomalous data points within a data set.	Numerical/TXT

MODEL REGISTRY

MODEL REGISTRY

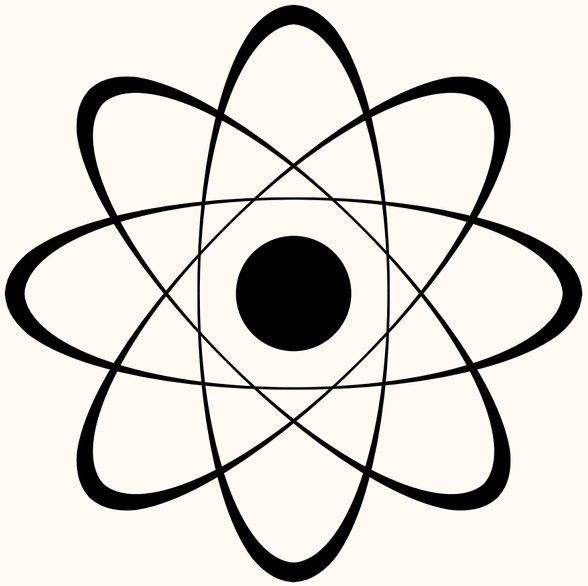
This is a central repository for your models. It becomes easier to track and have a catalog for your machine learning models.

MODEL REGISTRY



- You can register your trained machine learning models with the model registry.
- This allows you to catalog your models for production purposes.
- You can view the traceability and model lineage.
- You can deploy the model to production and also share the models with other users.

MODEL REGISTRY



- We first need to create a Model group that would have different versions of the model.
- You would create a model group for a particular business problem.
- All of the required models can then be part of the model group.

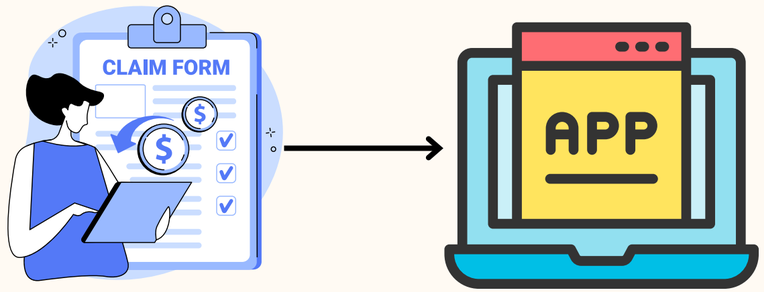
AWS MANAGED AI SERVICES

USING AI SERVICES ON AWS

AMAZON AI SERVICES

AWS has a lot of AI built in services that developers can incorporate in their applications. Instead of building the AI functionality from scratch , they can make use of existing Amazon services.

AMAZON AI SERVICES



Let's consider an application needs to have the capability to extract content from medical claim forms submitted by users.

- The medical forms would be physical in nature.
- An admin might scan and upload the medical claim forms.
- You need to develop code that would actually extract text based on the scanned document.
- And once you get the extracted text, then you need to segregate the different elements on the medical claim form.

AMAZON AI SERVICES

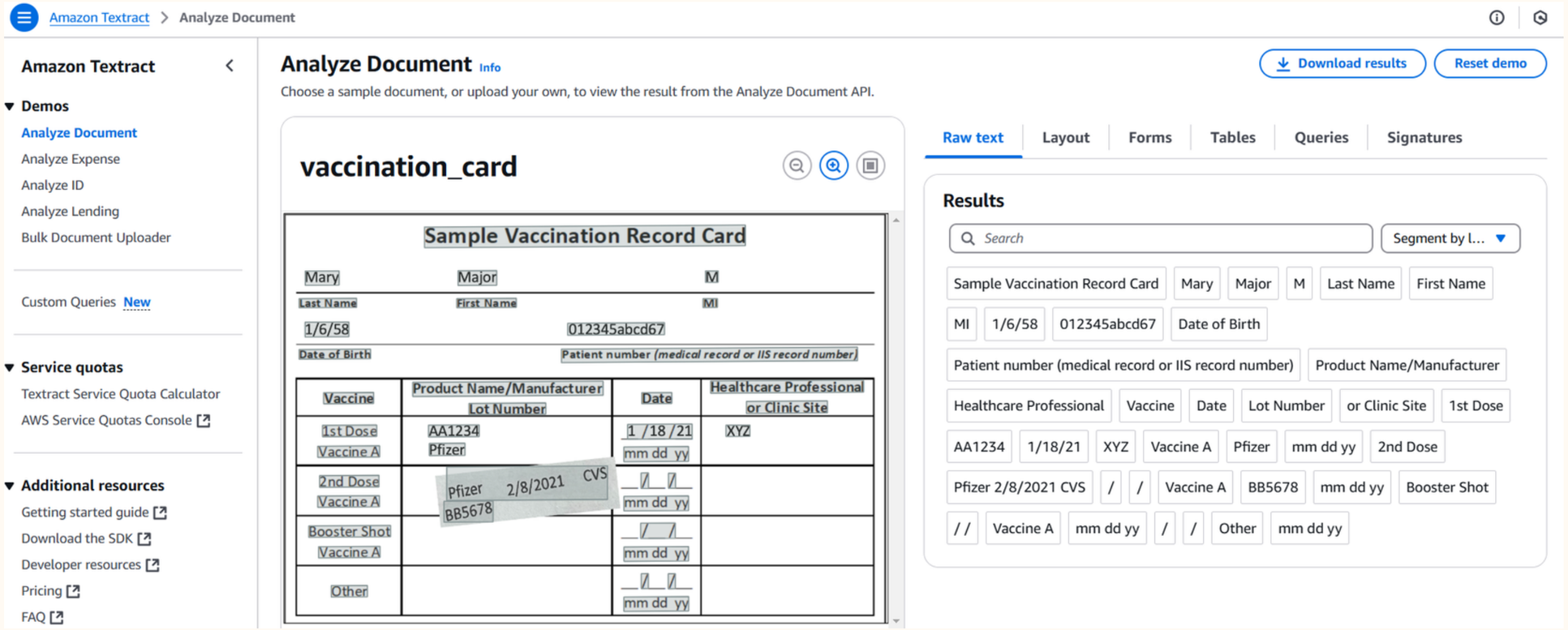
You need to develop code that would actually extract text based on the scanned document.

And once you get the extracted text, then you need to segregate the different elements on the medical claim form.

- It's not easy to develop code to perform the required functionality.
- And what if the requirement shifts from not only taking data from medical claim forms but from other different types of forms as well.

AMAZON AI SERVICES

We can leverage in-built services on Amazon.

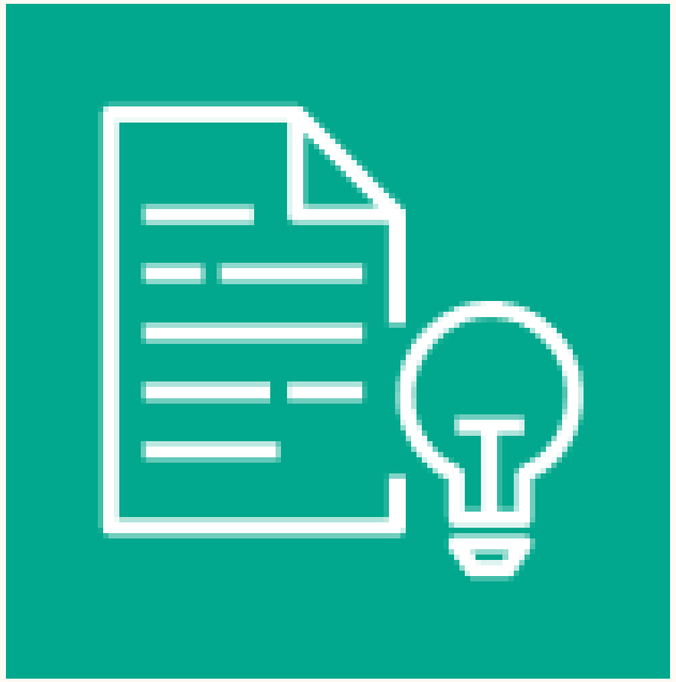


AMAZON COMPREHEND

AMAZON COMPREHEND

This service uses Natural Language Processing to extract insights from content. You can carry out a real -time analysis of your content or perform batch jobs.

AMAZON COMPREHEND



This service uses a pre-trained model that can be used to examine content within documents.

Entities - You can extract people, places, items and locations 01

Key phrases - You can extract key phrases. 02

PII - You can extract Personally Identifiable Information 03

Sentiment - The sentiment - positive, negative or neutral. 04

Language - The Language of the document. 05

AMAZON COMPREHEND

- We can also customize the use of Amazon Comprehend. And we don't need to have the skillset to know about machine learning to build the custom solution.
- We can create custom classification models that can be used to organize documents into our own set of categories.
- We can create a custom entity recognition model that can be used to analyze text for specific terms.

AMAZON TEXTRACT

AMAZON TEXTRACT

This service can be used to detect the typed and handwritten text within documents. It can also be used to extract text , forms and tables from within documents that contain structured data.

AMAZON TEXTRACT

- **Detecting Text**: Here we can get the lines and words of the underlying text. We can also detect the relationship between the lines and words of detected text.

- **Analyzing documents** You can extract data from forms and display them as key-value pairs. You can also extract table information. And the results can be returned in the form of JSON, CSV or a TXT-based file.

- **Analyzing invoices and receipts** We can also extract information from receipts and invoices.

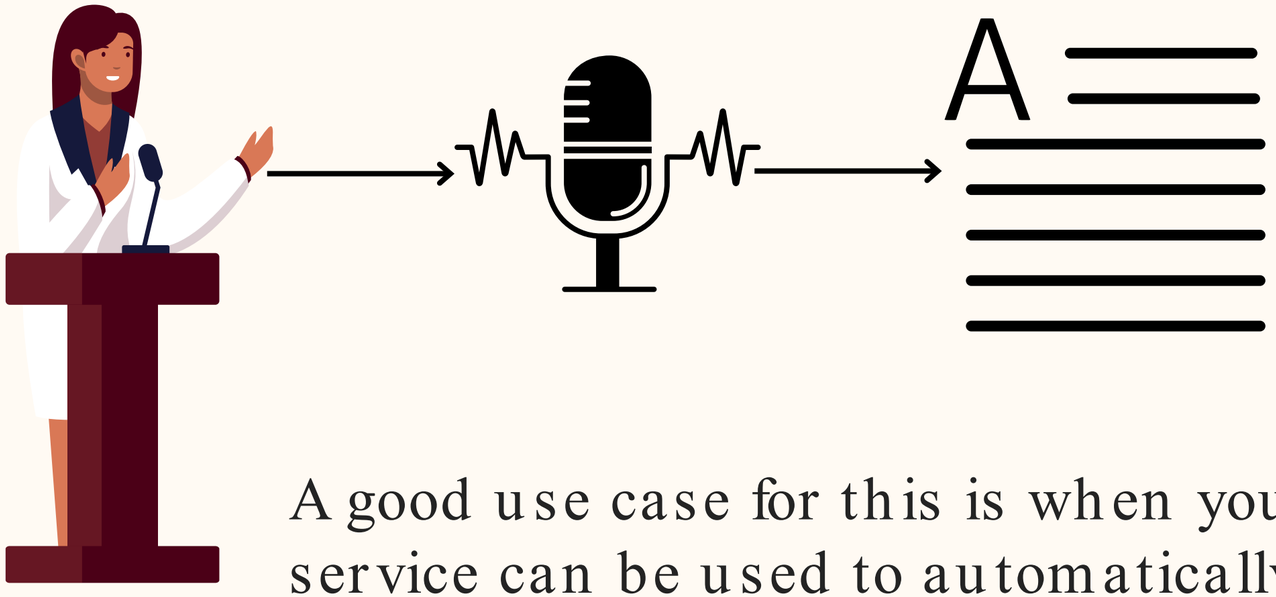
AMAZON TRANSCRIBE

AMAZON TRANSCRIBE

This service can be used to convert audio to text. It uses a machine learning model to achieve this.

AMAZON TRANSCRIBE

We can use this service to convert audio to text.



A good use case for this is when you want to transcribe speech. The service can be used to automatically convert audio to text.

AMAZON TRANSCRIBE

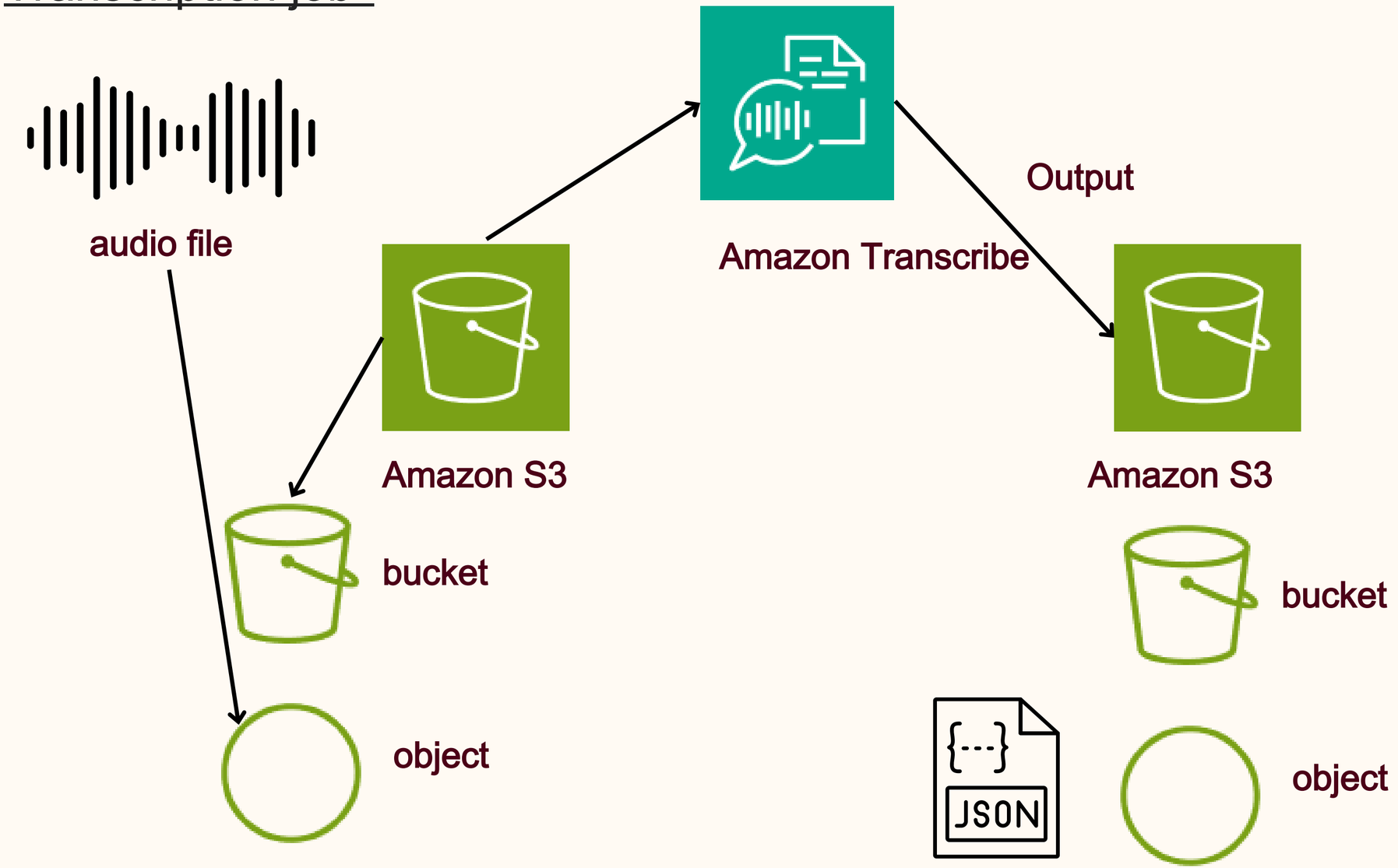
- We can perform real-time transcription. We can start streaming audio and get an immediate transcription when it comes to audio.
- We can also make the service detect the underlying language for the audio file.
- **Vocabulary filter**: We can provide a list of words that need to be modified in the output. For example, we might want to remove offensive words from the transcription output.
- **PII** - We can also identify and redact personal identifiable information - People names, email, Bank account details.

AMAZON TRANSCRIBE

- Supported audio formats for batch processing - MP3, MP4, WAV etc.
- There is support for both single and dual-channel audio.
- The output of the transcription will be in JSON.

AMAZON TRANSCRIBE

Transcription job



AMAZON REKOGNITION

AMAZON REKOGNITION

This service uses pre-trained computer vision based models that can be used to analyze images and videos. This service uses deep learning to analyze the images and videos.

AMAZON REKOGNITION

01

We can detect and classify objects, scenes and celebrities within images.

02

Detect and recognize printed and handwritten text within images.

03

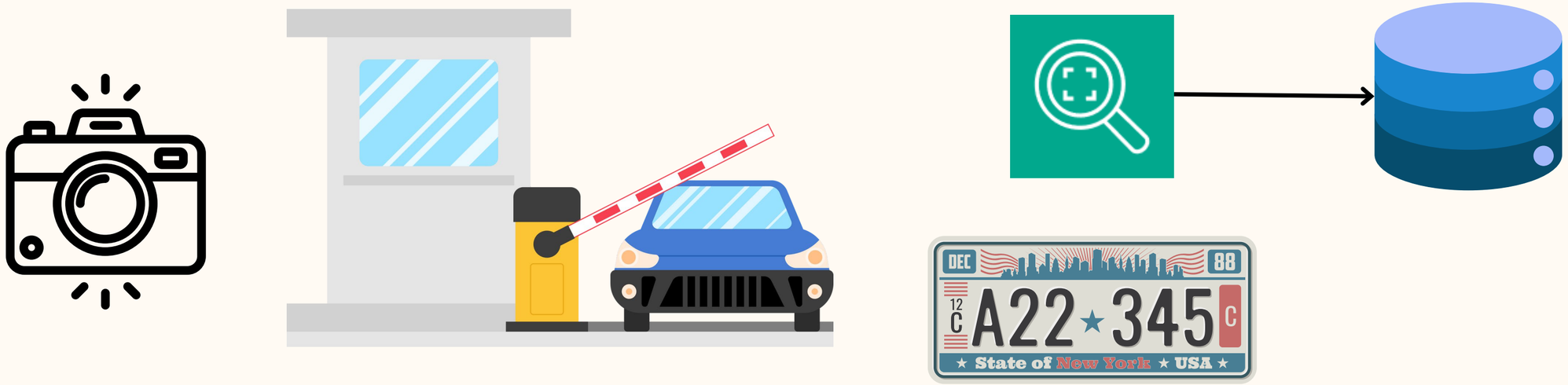
We can detect , analyze and compare faces.

04

We can detect and filter inappropriate and violent content.

AMAZON REKOGNITION

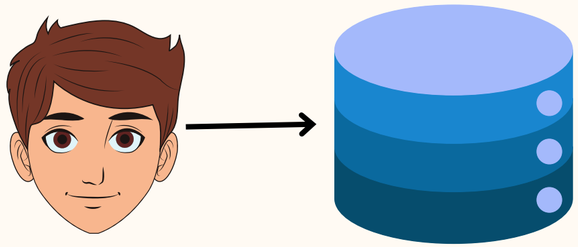
Detecting and storing car number plates



When vehicles pass through a parking gate, the camera can capture the image of the license plate. Amazon Rekognition can then extract the number plate and store in a data store.

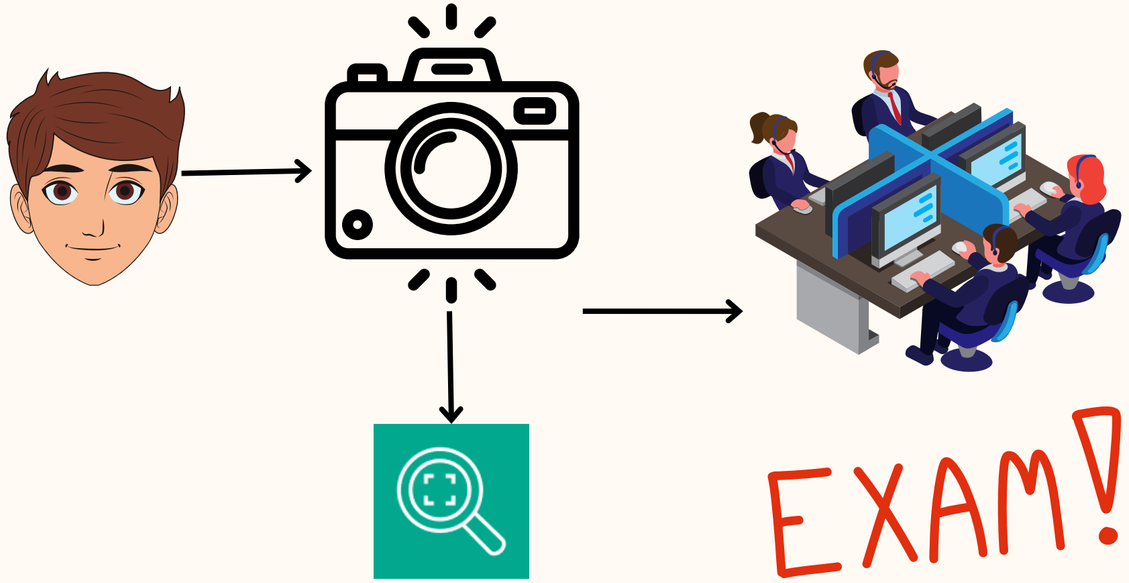
AMAZON REKOGNITION

Facial Recognition



When giving online exams, we first register ourselves onto a platform.

We can use a software on a mobile device to take a selfie and that is stored in a data store.



On the day of the exam, we need to enroll. At that time a picture is taken. The service can compare the picture with the one stored in the data store. This helps to verify the identity of the individual.

AMAZON POLLY

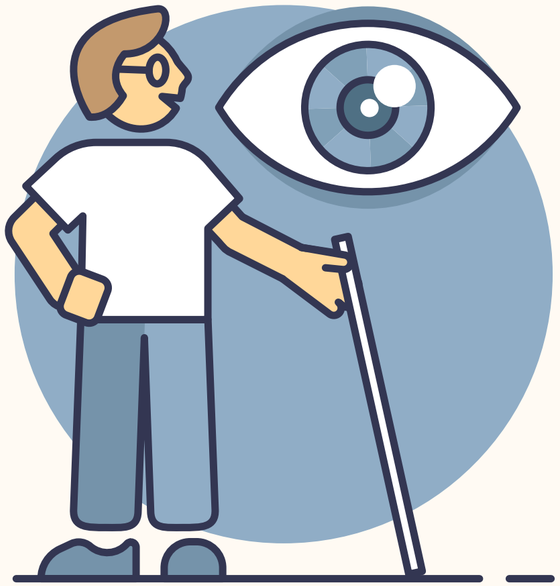
AMAZON POLLY

This service can be used to convert text to speech. This service supports multiple languages and many lifelike voices.

AMAZON POLLY

- You can submit text and plain -text to the service. Or provide the input in Speech Synthesis Markup Language (SSML).
- You can choose from a portfolio of voices available on AWS.
- The output audio file can be in a variety of file formats.

AMAZON POLLY



There are many use cases for this service. An example is applications used by the visually impaired. Amazon Polly can be used to generate audio for text on the screen of the devices.

AMAZON FORECAST

AMAZON FORECAST

This is a fully managed service that is based on machine learning algorithms. This can be used for forecasting or predicting values based on historical time series data.

AMAZON FORECAST

01

Importing your training data.

You first need to feed in your training data to Amazon Forecast. This is the historical data that is available. This data needs to be loaded in Amazon S3. Then you import your time series data to the service.

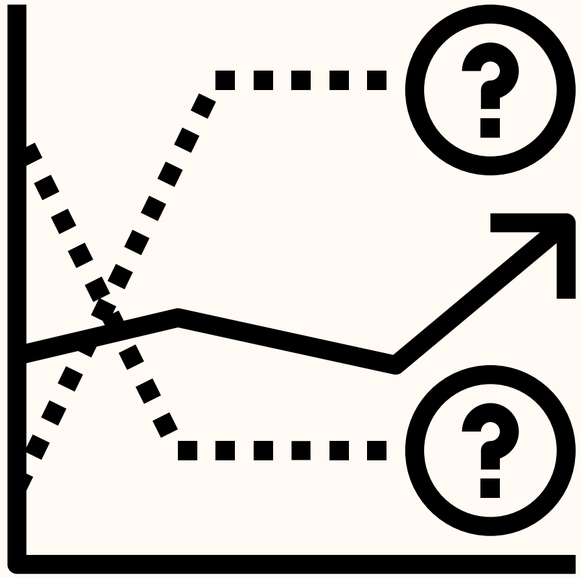
AMAZON FORECAST

02

Create a predictor

Here you create a predictor. This is used to generate forecasts based on the time series data. Amazon Forecast will then train the predictor based on the fed -in training data in the prior step.

AMAZON FORECAST



03

Create a forecast

Once you have the predictor in place, you can create a forecast. You can feed in your target data set and get predictions accordingly.

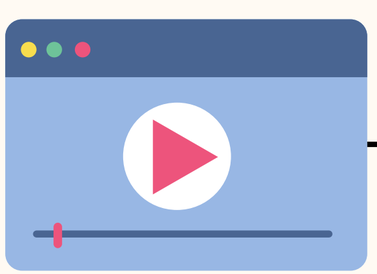
AMAZON TRANSLATE

AMAZON TRANSLATE

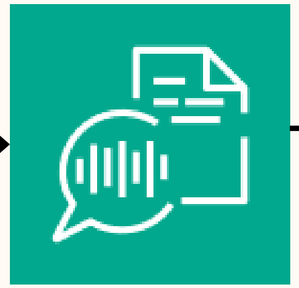
This is a fully managed service that is used to convert text from one language to another. This service again used machine learning to achieve this.

AMAZON TRANSLATE

Use case - Generating captions in different languages.

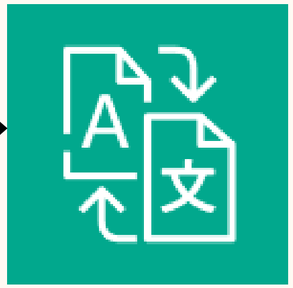


Step 1 : Extract audio from the video file.



Amazon Transcribe

Step 2 : Use the Amazon Transcribe service to convert audio to text



Amazon Translate

Step 3 : Use the Amazon Translate service to convert text from one language to another and use the result as captions for the video.

AMAZON TRANSLATE

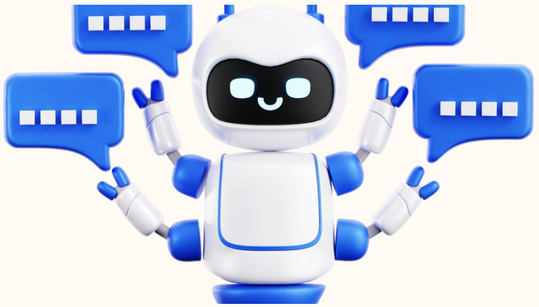
- Amazon Translate has support for a variety of languages when it comes to translation.
- You can provide input as text in UTF -8 format.
- Or having a file or collection of files in formats - .txt, .html. The output is given in the same format as the input file.
- You can also customize the translations - You can mask profane words in the output.

AMAZON LEX

AMAZON LEX

This service can be used to build conversational bots. The bots can be used to engage with users in a human -like way.

AMAZON LEX

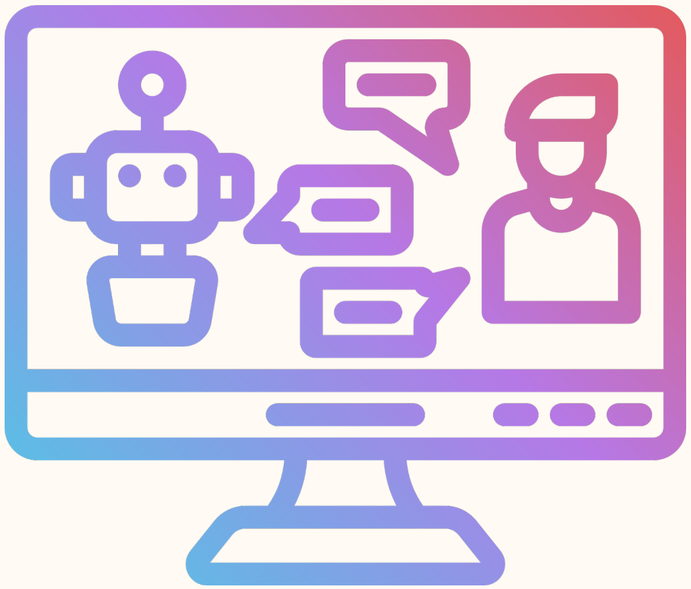


01

Step 1 : Create the bot and configure the intents.

The intents are what the user wants to achieve. In order to achieve this, the user will usher utterances and converse with the bot. We need to code responses for the utterances.

AMAZON LEX

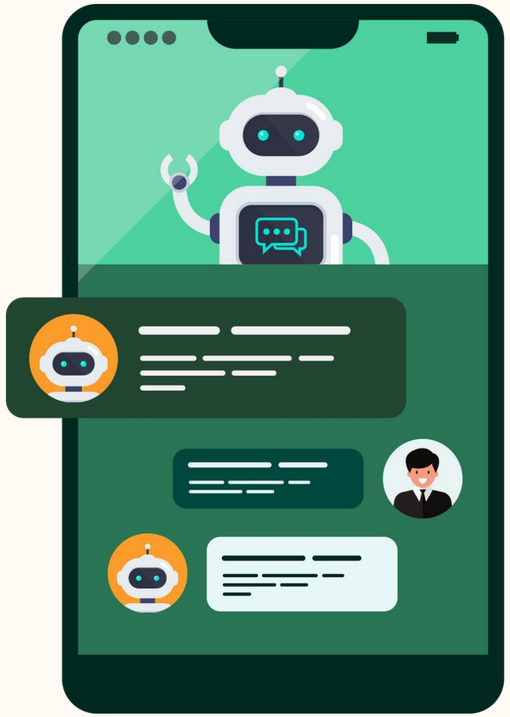


01

Step 2 : Test the bot.

Make sure the bot is working as expected.

AMAZON LEX



01

Step 3 : Publish and deploy

Publish and deploy the bot.

AMAZON PERSONALIZE

AMAZON PERSONALIZE

This service can be used to generate recommendations for users based on existing data. Here Machine Learning is used to generate the recommendations.

AMAZON PERSONALIZE

If you look at video streaming services, you might want to give your existing users suggestions on what to view next. Based on what they have currently seen, you might want to suggest on what to see next.

AMAZON PERSONALIZE

For an ecommerce site, based on the recent purchases or searches, we might suggest what the user could purchase or consider.

AMAZON PERSONALIZE

- First, we need to have an existing data set in place - For example for an ecommerce web site, we need to have purchases which were made in the past, correlated items etc.
- We import the dataset into the Amazon Personalize service.
- Based on the dataset we can start creating solutions.

AMAZON KENDRA

AMAZON KENDRA

This is a managed service used for information retrieval and search. It uses natural language processing and deep learning to provide efficient information retrieval and search capabilities.

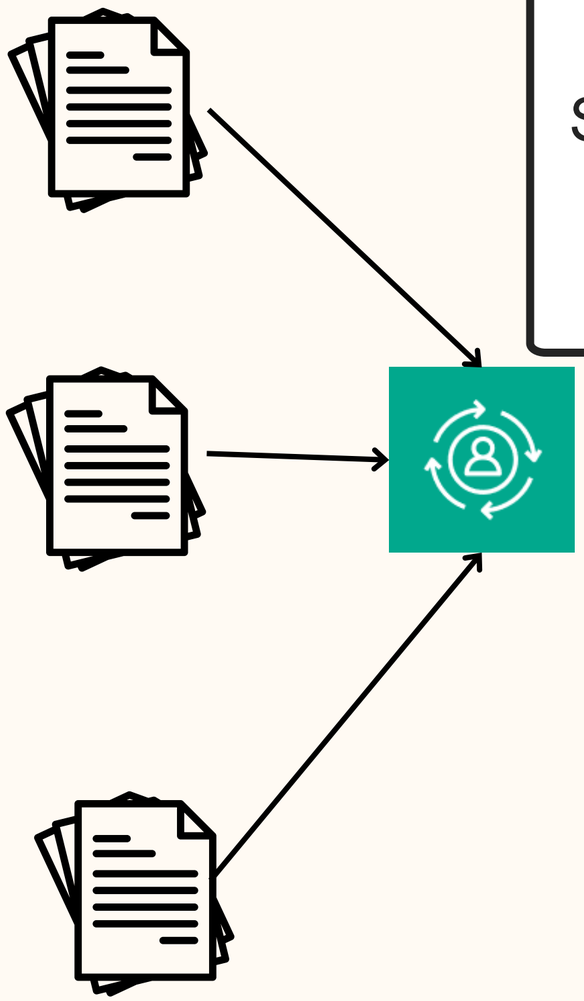
AMAZON KENDRA

You can have documents stored in Amazon S3 or external data sources such as Microsoft SharePoint or Salesforce. You can use the in -built connectors to crawl and index the documents stored in these various sources.

Microsoft Sharepoint

Amazon S3

Salesforce



AMAZON KENDRA

- Amazon Kendra supports documents in different formats - PDF, HTML, Word , PowerPoint etc.
- Amazon Kendra can extract the data in the documents and make them searchable in nature.
- Users can then query the indexes.

GENERATIVE AI

AMAZON BEDROCK

AMAZON BEDROCK

This is a managed services that allows you to work with the various foundation models. Models from Stability.ai, Hugging face and other companies are available for use.

AMAZON BEDROCK

- This is a fully managed service that allows you to make use of foundation models.
- You have one place where you can try out the different models - You can submit prompts and see how the responses get generated.
- You can customize an Amazon Bedrock foundation model with the use of training data.
- You can implement guardrails to implement safety features for your AI -based applications.

AMAZON BEDROCK

- You can create knowledge bases via the concept of RAG - Retrieval Augmented Generation.
- Here you can setup data sources from your internal data stores. And allow your foundation model to access these data sources.
- You can also automate tasks with the help of AI agents.

CHOOSING A MODEL

CHOOSING A MODEL

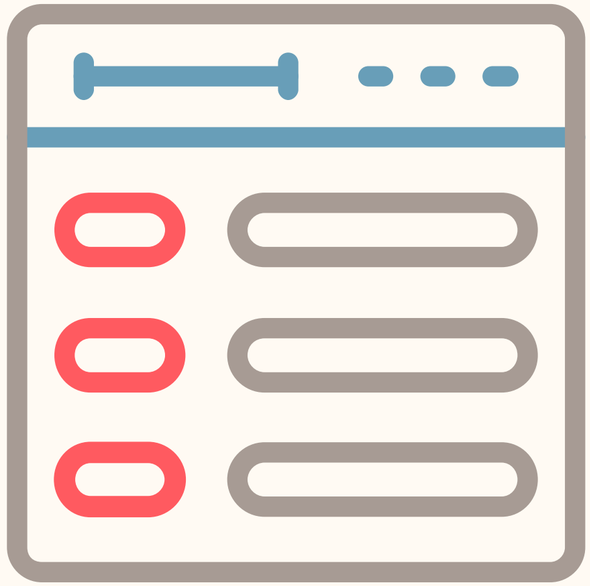
There are so many models to choose from. How do you know which model to select.

SELECTING A MODEL

Requirement

- Business requirement - What is the business problem, what is the requirement, what is the end goal.
- Is there a proper business case for using a foundation model.
- In the end using a model is costly. You need the infrastructure to support the model.

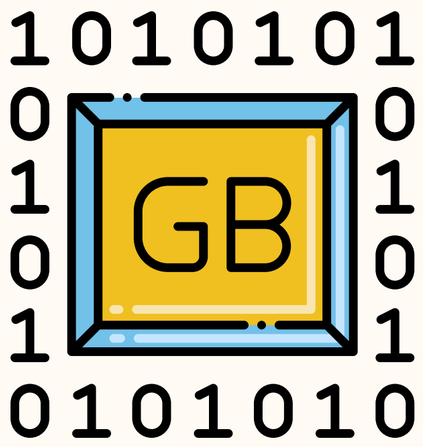
SELECTING A MODEL



Modality

- What is the modality of the solution.
- What is the input and output expected?
- Do you need a model that would provide text generation or do you need a solution that would generate images.
- Based on the modality you can limit the choice when it comes to foundation models.

SELECTING A MODEL



Size

- Once you have selected the modality, and you have narrowed down your choices, you need to look at the model size.
- Let's say for your foundation model you have narrowed down on GPT-4, Claude and Llama. Llama has models with 8 and 70 billion parameters.
- The larger number of parameters, the more accurate you can expect the model to be. Because the model can use the parameters to better detect the patterns in the underlying data when being trained.
- But then we need to consider the size of the model and the infrastructure required.

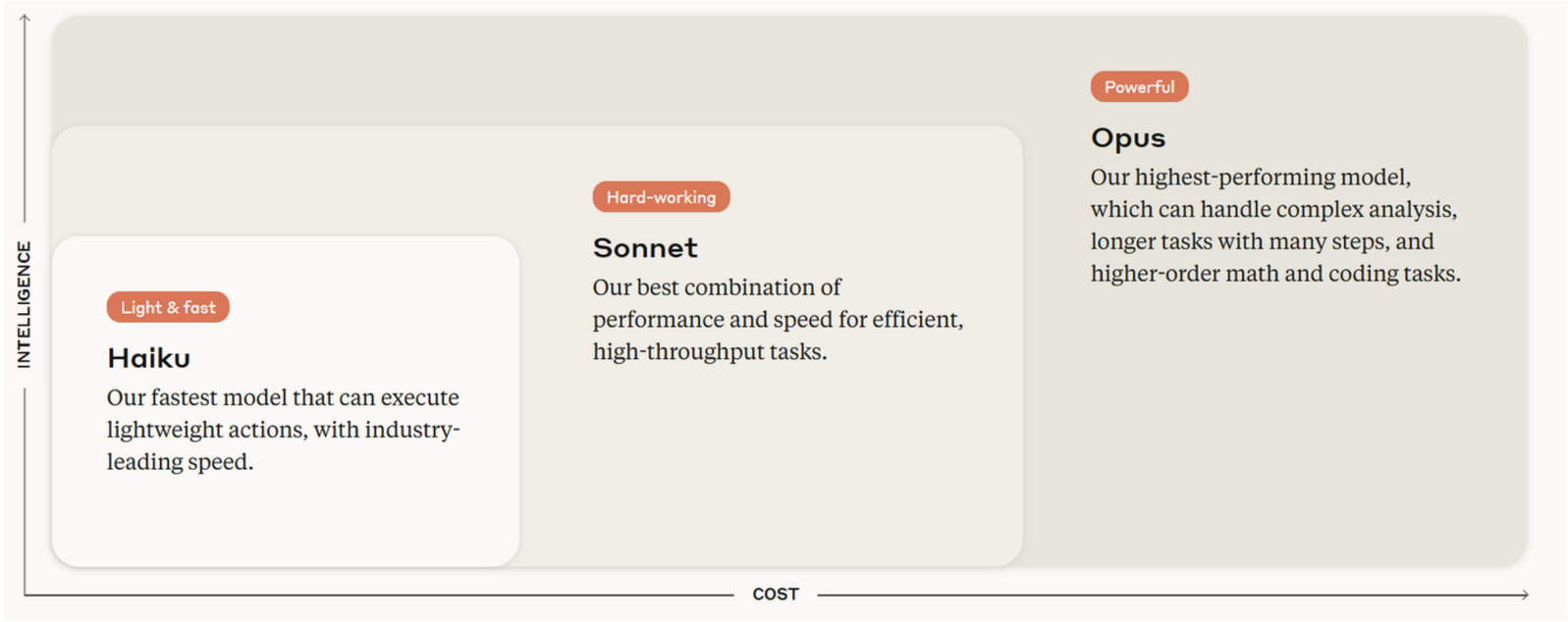
SELECTING A MODEL

Response time

- Next, we need to consider the response time.
- Each model can accept a certain number of input tokens. If it accepts more tokens and the size of the model is large, it probably has the capability to process more data at once.
- But then if more data needs to be processed and the model is large, then it would probably take more time to get the result. The response time needs to be considered.
- That is why a lot of companies will have different flavors for the model -

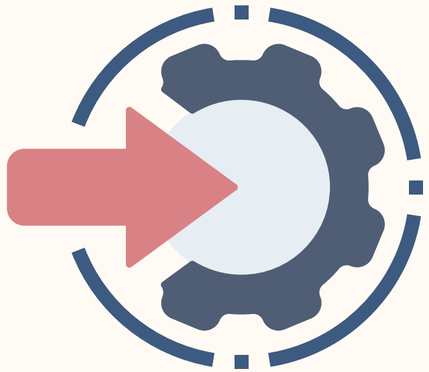
SELECTING A MODEL

Response time



Reference - https:/ / www.anthropic.com/ claude

SELECTING A MODEL



Context window

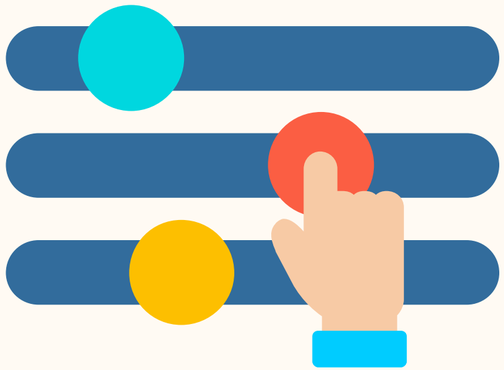
- Another factor to consider is the context window.
- This refers to the maximum number of tokens that can be given in one request.
- When it comes to text generation , this would include the input and the output tokens and any reasoning tokens if used.
- If we consider the GPT-4o model as an example, the context window is 128,000 tokens.

SELECTING A MODEL

Quality

- Next is the quality of the response.
- Does the responses have any bias, any hallucination. Are the results accurate.
- If the model is being used for critical applications, all of these factors need to be considered.

SELECTING A MODEL



Fine tune

- Is it possible to fine tune the model to make it better.
- Can we extend the functionality of the model to include the companies' data.
- Can the model work with additional data and provide responses based on additional data sources.

SELECTING A MODEL

Price

- The next and important factor to consider is the price.
- With better performing models , models that can take in a lot of data, the price might be higher. It's all about price over performance.

SELECTING A MODEL

Amazon Titan

Region:

US East (Ohio)

On-Demand and Batch pricing for text models

Amazon Titan models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Amazon Titan Text Embeddings V2	$0.00002	N/A	N/A	N/A

Pricing for Creative Content Generation models

On-Demand pricing for Image Generator Model

Region:

US East (N. Virginia)

Amazon Nova models	Image resolution	Price per image generated for Standard quality	Price per image generated for Premium quality
Amazon Nova Canvas	up to 1024 x 1024	$0.04	$0.06
Amazon Nova Canvas	up to 2048 x 2048	$0.06	$0.08

SELECTING A MODEL

Anthropic

On-Demand and Batch pricing

Region: US East (N. Virginia) and US West (Oregon)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)	Price per 1,000 input tokens (cache write)	Price per 1,000 input tokens (cache read)
Claude 3.5 Sonnet**	$0.003	$0.015	$0.0015	$0.0075	$0.00375	$0.0003
Claude 3.5 Haiku	$0.0008	$0.004	$0.0005	$0.0025	$0.001	$0.00008
Claude 3 Opus*	$0.015	$0.075	$0.0075	$0.0375	NA	NA

Stability AI

On-Demand pricing

Stability AI model	Price per generated image
Stable Diffusion 3.5 Large	$0.08
Stable Image Core	$0.04
Stable Diffusion 3 Large	$0.08
Stable Image Ultra	$0.14

SELECTING A MODEL

Anthropic

On-Demand and Batch pricing

Region: US East (N. Virginia) and US West (Oregon)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)	Price per 1,000 input tokens (cache write)	Price per 1,000 input tokens (cache read)
Claude 3.5 Sonnet**	$0.003	$0.015	$0.0015	$0.0075	$0.00375	$0.0003
Claude 3.5 Haiku	$0.0008	$0.004	$0.0005	$0.0025	$0.001	$0.00008
Claude 3 Opus*	$0.015	$0.075	$0.0075	$0.0375	NA	NA

Stability AI

On-Demand pricing

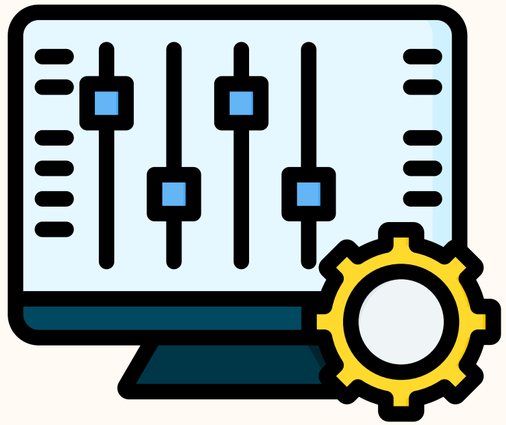
Stability AI model	Price per generated image
Stable Diffusion 3.5 Large	$0.08
Stable Image Core	$0.04
Stable Diffusion 3 Large	$0.08
Stable Image Ultra	$0.14

INFERENCE PARAMETERS

INFERENCE PARAMETERS

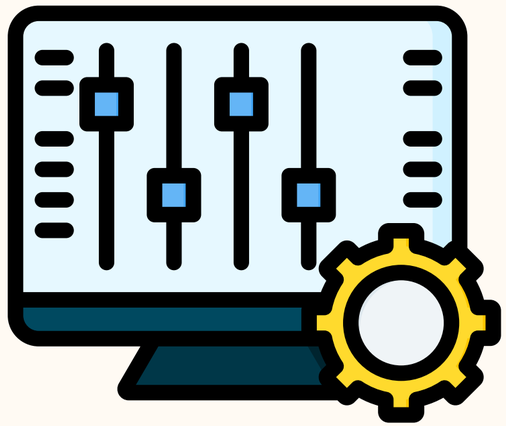
We can tweak settings known as inference parameters. This helps configure the responses returned by the model.

INFERENCE PARAMETERS



- **Prompt-** This is the input given to the model.
- **Inference parameters** This can be used to influence the model response.
- **Token-** A model uses this as a unit of input. It may be a word or part of a word.

INFERENCE PARAMETERS



GPT-4o & GPT-4o mini | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

We are hear to learn about Amazon Bedrock

Clear | Show example

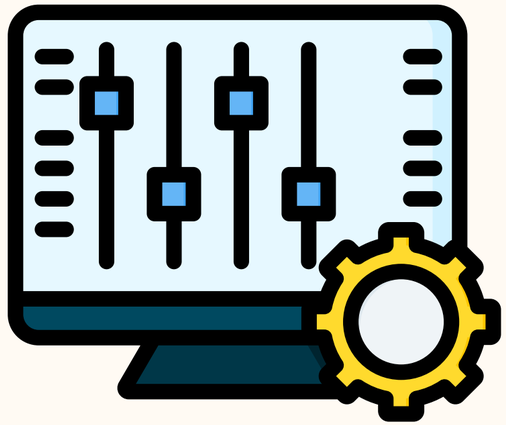
Tokens 9

Characters 41

We are hear to learn about Amazon Bedrock

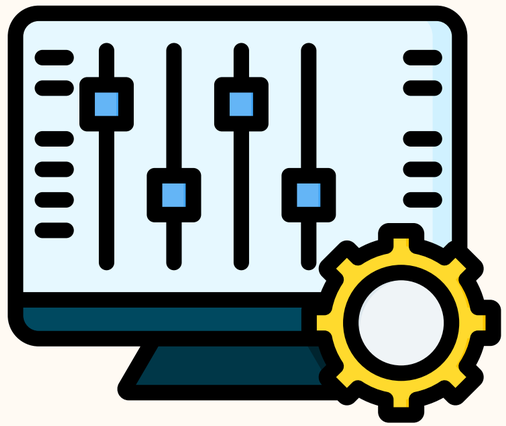
Reference -
https://platform.openai.com/tokenizer

INFERENCE PARAMETERS



- The inference parameters can influence the next token that is sent by the model in the response.
- **Temperature** – This effects the probability distribution of the outcome.
- **Lower temperature** This leads to high probability outputs. It gives more deterministic responses.
- **Higher temperature** This leads to lower probability outputs. It gives more random responses.

INFERENCE PARAMETERS



- **Top K**- This number determines the number of candidates the model considers for the output tokens.
- A lower value will make the model consider less tokens as options for the next token in the output.
- **Response Length**- This is if you want to limit the number of tokens in the response.
- **Stop sequence**- You can specify a sequence. The model will stop generating when it encounters the sequence.

PROMPT ENGINEERING

PROMPT ENGINEERING

This is the process of creating instructions that are sent to a Gen -AI model in an attempt to get the best possible output.

PROMPT ENGINEERING

- There are some simple rules when it comes to designing prompts.
- Remember that you want to extract a desirable answer from the model.
- Hence the first and most important step is to be clear in your instructions to the model. Don't be vague, otherwise the model will not clearly understand what exactly you want to achieve.

PROMPT ENGINEERING

- When it comes to a prompt, it can consist of several aspects.
- First is the task you want the LLM to perform. Maybe you want to summarize text or generate code.
- You could also tell the model that you want the response in a particular format. Hence you can provide additional instructions to the model.
- You could also give some examples on how the model can generate an answer.

PROMPT ENGINEERING

- **Zero-shot prompting**- Here we don't provide any examples to the model.
- We basically give an instruction to the model to get an outcome.
- **Few-shot prompting**- Here we provide some examples to the model. We want to steer the model in a particular direction to provide us with the required outcome.

PROMPT ENGINEERING

- **Chain-of-thought prompting** - This can be used to solve complex problems.
- Here we provide instructions which can be done via few - shot prompting on how to tackle a problem.
- We want to introduce a chain of thought to the model on how it should go about solving a problem.

PROMPT ENGINEERING

- **Prompt templates** - This is a prompt that contains variables. We can replace the values in the variables with different data for the prompts.

PROMPT ENGINEERING

- **Prompt persona** Here you can provide a system prompt.
- You want the model to adapt a persona and then respond to prompts accordingly.

AMAZON BEDROCK GUARDRAILS

AMAZON BEDROCK GUARDRAILS

This can help add safeguards to your generative AI-based applications. In addition to using Amazon bedrock with the various models, you can safety measures with the help of Guardrails.

AMAZON BEDROCK GUARDRAILS

- This feature can be used to filter harmful content in both the inputs and outputs given by models.
- For example, if the models are used to power chatbots, we can filter any sort of malicious content. And avoid harmful input reaching the model.
- We can redact responses that contain any sort of personal identifiable information .
- There are different types of policies that can be put into place with Guardrails .

AMAZON BEDROCK GUARDRAILS

- **Content filters** - This can be used to block inputs prompts or responses that contain harmful content.
- **Denied topics** - These are a list of topics that should be blocked.
- **Sensitive information filters** - This can be used to block or mask personal identifiable information .
- **Contextual grounding check** - This can be used to detect and filter hallucinations .
- **Image content filter** - This can help detect and filter inappropriate or toxic image content.

AMAZON BEDROCK GUARDRAILS

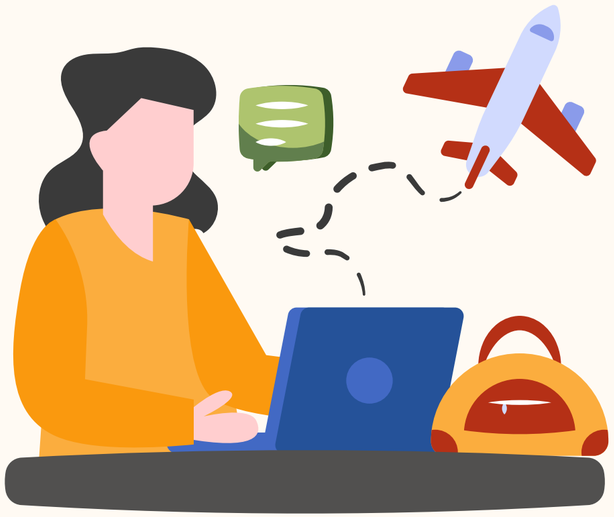
- You can create multiple guardrails with each guardrail having specific policies defined.
- Once you want to deploy the guardrail to production, after testing, just create a version of the guardrail .
- You can then invoke the guardrail via an API call from your application .

AMAZON BEDROCK AGENTS

AMAZON BEDROCK AGENTS

We can build agents that can orchestrate actions or a flow based on the user input. This helps to extend the value of a foundational model for an application.

AMAZON BEDROCK AGENTS



- Amazon Bedrock agents can orchestrate actions based on the interaction of the foundation model and the user in addition to making use of data sources and external applications .
- For example, you can make a complete workflow that could making flight bookings for customers .
- First the foundation model could help make the itinerary based on the user's input and requirement .
- Then the agent could call external applications and make the bookings accordingly .

AMAZON BEDROCK AGENTS

- When configuring the agent, you can also create a knowledge base in Amazon Bedrock to help facilitate the actions taken by the agent.
- For the agent you will choose a foundation model which will primarily be used to interact with the user via prompts and provide responses.
- **Action Groups** - You define actions that the agent needs to perform .
- You define a schema that would be used to extract parameter details based on the user interactions .

SECURITY AND MONITORING ON AWS

IAM USERS AND GROUPS

IAM USERS AND GROUPS

We can create principals on AWS that we can give access to resources accordingly. These principals can be AWS IAM Users.

AWS IAM USERS AND GROUPS

- **IAM User**- This is an entity that can represent a human user or workforce to whom we can give permission to resources.
- **Console password**- Here we can create a password for the user. The user can use this to authenticate and use the AWS console.
- **AWS Access Keys** This can be used to make programmatic calls to AWS,
- An IAM user can only be associated with one AWS account.
- By default, a new IAM user is not given any permissions.

AWS IAM USERS AND GROUPS

- **IAM Groups** - This is a collection of IAM users.

- If you want to manage permissions for multiple users at a time, you can place them as part of a group and give permissions to the group accordingly.

- An IAM user can belong to multiple user groups.

- Nesting of user groups is not possible; you can't have one group that is part of another group.

AWS CLOUDTRAIL

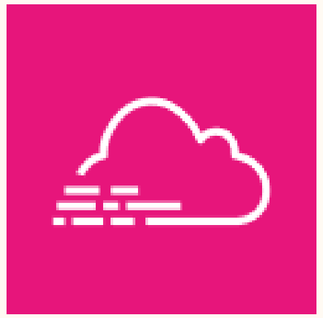
AWS CLOUDTRAIL

This service helps companies from an auditing and compliance perspective. This service records the actions taken by users, roles or AWS services.

AWS CLOUDTRAIL

- Event action taken on your AWS account gets recorded as an event in AWS CloudTrail.
- Whether the actions are taken via the AWS Console, or via programming calls, all of them get recorded in AWS CloudTrail.
- Currently there are four types of events that get recorded - Management, Data , Network Activity and Insight events.
- The CloudTrail service by default saves each event for a duration of 90 days. If you need the events for a longer duration of time you can create a trail that can persist the events to an S3 bucket.

AWS CLOUDTRAIL



- When persisting logs to Amazon S3, the logs can grow over time. You can save on storage costs by looking at changing the storage class for the objects in S3.
- **Standard S3**- This is the default object storage class in Amazon S3.
- **S3 Standard-Infrequent Access** This is ideal for data that needs high durability and throughput. But is not accessed that frequently.
- **S3 Intelligent-Tiering**- This service can automatically transition your objects to different classes based on the access. This can help to reduce the costs of objects at a granular level.

AMAZON MACIE

AMAZON MACIE

This is a data security service that can help discover sensitive data. It uses Machine Learning to identify potential data security risks.

AMAZON MACIE

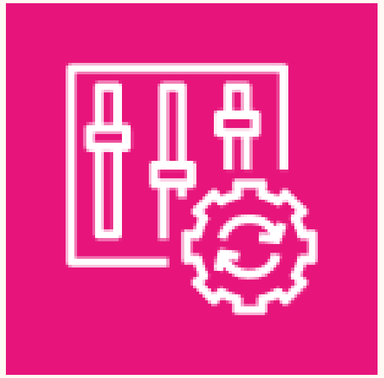
- Amazon S3 is used for storing a lot of data. There can be a lot of sensitive data stored within Amazon S3.
- To help maintain the security posture of the data within Amazon S3, we can make use of the Amazon Macie service.
- Here Amazon can continuously evaluate the permissions given to your S3 bucket. If it detects that the bucket containing your data has become publicly accessible, it can notify you accordingly.
- It can also detect sensitive data within your Amazon S3 buckets.

AWS CONFIG

AWS CONFIG

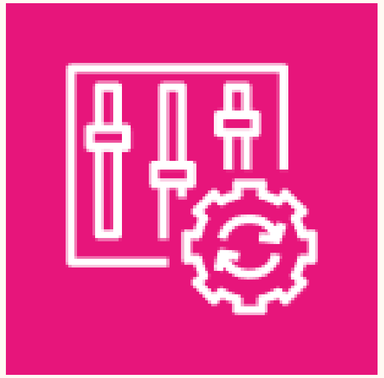
This helps to track the configuration of your resources. With the help of AWS Config rules , you can also check for the configuration of your AWS resources.

AWS CONFIG



- For you AWS resources, there will always be configuration changes made to the AWS resources over time.
- You can keep track of the configuration changes of resources via the use of AWS Config.
- In order to start using AWS Config, the service first needs to find all resources in your AWS account. It then generates configuration items for the changes made to your AWS resources.
- You can also use AWS Config rules to check for changes made to your AWS resources.

AWS CONFIG



- For example, you might setup a secure environment in which resources in a VPC communicate with Amazon Bedrock via the use of AWS PrivateLink .
- So, you would have made configuration changes to your resources to create this secure environment.
- You can create AWS Config Rules to trigger alerts when resource configuration cause the environment to deviate from the organization's compliance rules for the environment.

AWS AUDIT MANAGER

AWS AUDIT MANAGER

This helps to track the configuration of your resources. With the help of AWS Config rules , you can also check for the configuration of your AWS resources.

AWS AUDIT MANAGER

- This service helps to continually audit the usage of AWS and helps to manage risk and compliance with industry standards.
- There are pre-built frameworks that help to automate assessments when it comes to certain standards and regulations.
- You create assessments, the assessments then collect data for the AWS accounts.
- The data is collected continually till the assessment is active.
- You can use AWS Audit Manager to check if Generative AI applications hosted via the use of Amazon Bedrock are going against AWS best practices.

AWS AUDIT MANAGER

- The AWS Audit Manager Framework library has a framework when it comes to generative AI best practices.
- These practices focus on key areas such as governance, data security, data privacy etc.
- These are already mapped to Amazon Bedrock resources.
- Hence you can use this to create assessments for your applications running on Amazon Bedrock.

AWS ARTIFACT

AWS ARTIFACT

This service provides on -demand downloads when it comes to various security and compliance documents.

AWS ARTIFACT

- Here you can get reports which relates to different standards such as ISO (International Organization for Standardization), PCI (Payment Card Industry)
- You can also download the security and compliance documents for independent software vendors who sell their products on the AWS Marketplace.
- If auditors wants to get security and compliance information for whether AWS services follow industry standards, you can download the compliance documents using the AWS Artifact service.

AWS TRUSTED ADVISOR

AWS TRUSTED ADVISOR

This service can check the resources deployed in your environment and give you recommendations on how improve upon aspects such as cost, security and performance.

AWS TRUSTED ADVISOR

- For the Basic or Developer Support plan, you get access to five checks in the security category and all checks in the Service Limits category.
- You need to upgrade to the Business or Enterprise plans; you get access to all checks within AWS Trusted Advisor.
- You get access to checks in the following category
- **Cost Optimization** - Here you get recommendations on how you can save on costs. It will look at the deployment of your current resources and provide recommendations on potential cost savings.

AWS TRUSTED ADVISOR

- **Performance**- Here you get recommendations on how to improve on speed and responsiveness of applications to deployed to your AWS environment.
- **Security**- How you can make your AWS resources more secure.
- **Fault tolerance**- How you can increase the resiliency and redundancy of your solutions deployed to your AWS environment.
- **Service Limits**- Whether your account is reaching service limits for your AWS services.
- **Operational Excellence** How you can increase the operational efficiency of your environment.