



# KodeKloud

© Copyright KodeKloud

Visit [www.kodekloud.com](http://www.kodekloud.com) to learn more.



## Course Introduction

© Copyright KodeKloud

Welcome everyone let's talk a little bit about the structure of the certified AWD certified AI practitioner certification course

## Objectives

- 01 Fundamentals of AI and Machine Learning
- 02 Fundamentals of Generative AI
- 03 How to use Foundation Models with AI applications
- 04 Guidelines for responsible AI
- 05 Security, Compliance, and Governance for AI solutions
- 06 Preparing you for the AWS Certified AI Practitioner exam



This course by the way is designed to teach you the fundamentals of AI the fundamentals of generative AI specifically some of the foundational models and a lot of guidelines security and compliance for AI solutions the goal here is to prepare you for the aws certified AI practitioner exam

# Who Is It Designed for?

01



AWS Shared  
Responsibility Model

02



AWS Identity and  
Access Management

03



AWS Global  
Infrastructure

04



AWS Pricing models  
(e.g., usage and  
capacity)

Now unlike other foundational exams of which there's really only one other this exam is actually a little bit more advanced than say your traditional like cloud practitioner exam and oddly enough you actually need to have cloud practitioner or some facets of cloud practitioner to really be able to take the AI practitioner even though it's a foundational exam it seems to be slightly more complicated than just your traditional cloud practitioner exam because you do actually have to know the AWS care responsibility model you have to understand how security works in AWS you have to understand a little bit about AWS global infrastructure and you have to understand pricing all of this you actually get from the AWS

certified cloud practitioner exam which is the other foundational exam so it's kind of recommended that you actually have that exam or at least equivalent experience with a WS before you take this one.

# AWS Cloud Practitioner

AWS, Cloud, DevOps

© Copyright KodeKloud

So this course was designed with people who have some foundational knowledge of AWS because we're not going to cover things like pricing models or availability zones we assume that you already know that

# Main Content



5 Domains



7 Sections



Pre- and Post-  
Practice Exams



Demos

A visual representation of the course structure, showing the flow from domains to sections to various learning activities

© Copyright KodeKloud

Now this course outline basically there's five content domains that make up the core of this course we obviously have an opening section and a closing section but there's five content domains so 7 sections total we also offer like we traditionally do a pre and a post kind of assessment exam so you can find out whether you even need to take this of course and to see how much you've grown after you finish we also have demos we'll be throwing in some games but the main idea here is this course follows the structure of a lot of our other AWS courses with you know 5 content domains

## Introduction

Why take this course?

What are we covering?

Who is this course for?

How to prepare/study?

Are you ready to start?

Pre-assessment

## Closing

What did we study?

Are you ready for the exam?

What's next?

Resources for ongoing learning

So what that means is that Section 1 is gonna answer the normal questions like what are we covering who is this course for like the one you're taking right now and it also includes the pre assessment and a few little details about like how to sign up for the exam and what to expect we also have a closing section that confirms kind of what you study Are you ready for the exam and some further resources for study because this is a foundational exam so there's other

# Pre- and Post-Assessments



**65**

Questions

B E F O R E <



**100**

Minutes

> A F T E R

© Copyright KodeKloud

Our pre and post assessments match what's inside of the exam meaning the exam is 65 questions and it's got 100 minutes to do that so we've created the 65 question pre and post assessment that allows you to train and assess whether or not you're actually ready to take the actual exam

# 5 Content Sections

01



Fundamentals of AI  
and ML

02



Fundamentals of  
Generative AI

03



Applications of  
Foundation Models

04



Guidelines for  
Responsible AI

05



Security, Compliance,  
and Governance for  
AI Solutions

And then just like we reviewed as far as the objectives there are basically 5 content sections these content sections are the fundamentals of AI the fundamentals of generative AI all the things that you see listed here we will be having quizzes in those sections and games as well to help reinforce the learning

# An Even More Important Approach Than Before



© Copyright KodeKloud

Now we talked about this in almost every AWS course we have which is basically be patient be consistent play with a WS right and then make sure your time is focused and protected

1 you gotta be patient with yourself this is a marathon not a Sprint and while you can move swiftly you're gonna have to move and pace yourself because it is a decent amount of material there's usually an estimated of 40 hours for the average learner to actually take this in from scratch if you don't have prior AWS experience

2 consistency is key meaning a little bit every day is going to be extremely important this is better to do 15 minutes in a single session than it is to do a four hour session say on like a Saturday

three if possible do the labs we have playgrounds we have labs you have the opportunity to get your hands on a WS particularly if you're struggling with a concept

four and last probably the most important protect your time if you've got 15 minutes free then study right do the mock exam read the material stay focused it's going to require consistent focus to drive this new information inside of you and I could argue personally that focus is probably your only superpower so keep going

## Summary

- 01 This course is designed for people who have Cloud Practitioner-level knowledge or equivalent experience with AWS.
- 02 It has five content sections aligned with each Domain in the AWS Exam Guide.
- 03 It has pre- and post-assessments, with quizzes, demos, and games in between.
- 04 Learners must create a schedule and be consistent in their studies to prepare for the exam.





KodeKloud

# Amazon Web Services (AWS) Certified AI Practitioner Certification

---

Why Take This Course?

© Copyright KodeKloud

Welcome everyone and let's talk about why the certified AI practitioner certification is going to be so important not just now but in the future

# Industry Demand



AI and GenAI are going to be significant technologies not just in cloud computing but in almost everything we do in IT.

© Copyright KodeKloud

1st I just want to say that if you are not aware AI and generative AI in particular are pretty much taking the world by storm particularly IT AWS is one of the first cloud providers to provide an AI certification along with Azure and along with GCP so this is significant because industry demand for AI and Gen. AI skills are now a requirement much like knowledge of Kubernetes and cloud skills are a requirement for an operations role

## Foundations

**This is a Foundational exam, and it will lay the groundwork for future AI-enabled certifications including Data Engineering and Machine Learning.**

© Copyright KodeKloud

This is also going to set the foundations for you to take further exams this is the foundational knowledge not just for AI exams but for data engineering and machine learning exams that are going to be emerging on the AWS platform so knowing this material and this information is going to help you take those very AI focused exams

# Compensation



KodeKloud

This exam has become part of a portfolio of emerging AI technologies that you can signal your growing skill and interest in.

© Copyright KodeKloud

Now while there are no promises here at all about what compensation you could achieve because this is a foundational exam it is one of the only ways currently in 2024 to signal that you have studied AI on AWS and that you have an understanding of the fundamentals of what services can support the guy on a WS there is no other exam currently as of October 2024 that currently will send that signal and so this is one of the best ways to send that signal and therefore you will be compensated for it

## Hiring Advantage

**Obtaining an AWS Certified AI Practitioner certification is still a significant hiring advantage compared to someone who doesn't have it.**

© Copyright KodeKloud

This is also a distinct hiring advantage so not only could you possibly be more valuable to a company and therefore be compensated better but if it comes down to people with the same skills and you have some AI in your portfolio including the specific certification then when it comes down to a comparative analysis if you have AI and the other person doesn't you win

# AWS Certification Roadmap

## Foundational

Knowledge-based certification for foundational understanding of AWS Cloud

No prior experience needed



## Associate

Role-based certifications that showcase your knowledge and skills in AWS and build your credibility as an AWS Cloud professional

Prior cloud and/or strong on-premises IT experience recommended



## Professional

Role-based certifications that validate advanced skills and knowledge required to design secure, optimized, and modernized applications and to automate processes on AWS

Two years of prior AWS Cloud experience recommended



## Specialty

Dive deeper and position yourself as a trusted advisor to your stakeholders and/or customers in these strategic areas

Refer to the exam guides on the exam pages for recommended experience



Now it would be totally disingenuous to at least not put you in context that this foundational exam it is currently in beta but will quickly move out of beta and that this is the start of a whole slew of other exams that you can take



## Summary

01

This course is an essential starting point for those wanting to start their AI journey on AWS.

02

By the end of this course, you will be ready to take the AWS Certified AI Practitioner certification exam.

03

Some reasons to take the course: Emerging AI dominance in IT, sending recency signals to employers, possible increase in compensation for skills, etc.

04

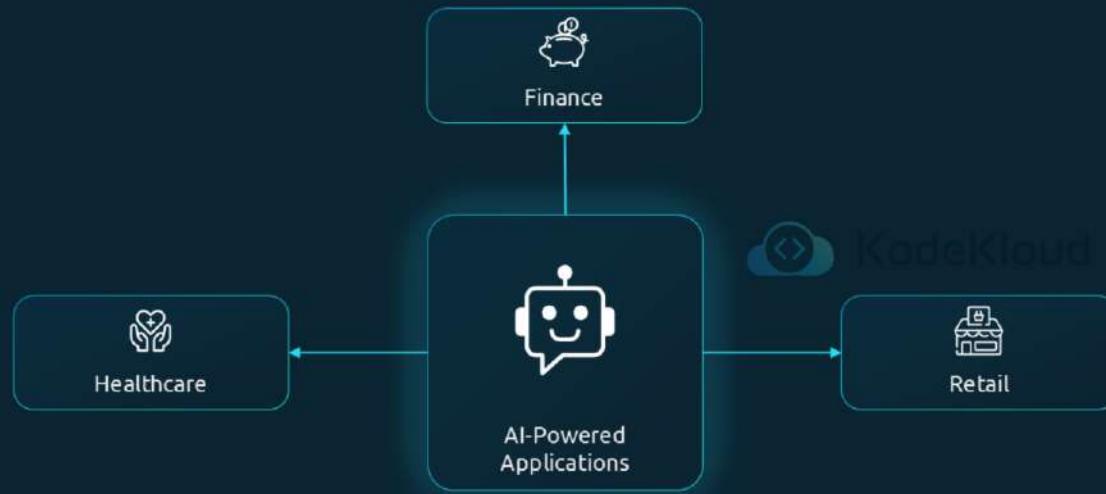
Show your employers that you know how to operate the AWS platform, specifically for AI on AWS.





# Basic AI Concepts and Terminologies

# Why Artificial Intelligence (AI) Matters



# Why Artificial Intelligence (AI) Matters

01



Improving efficiency

02



Reducing  
operational costs

03



Enhancing  
decision-making

AI is revolutionizing multiple industries by improving efficiency, reducing operational costs, and enhancing decision-making.

# Why Artificial Intelligence (AI) Matters



Organizations



Customer service



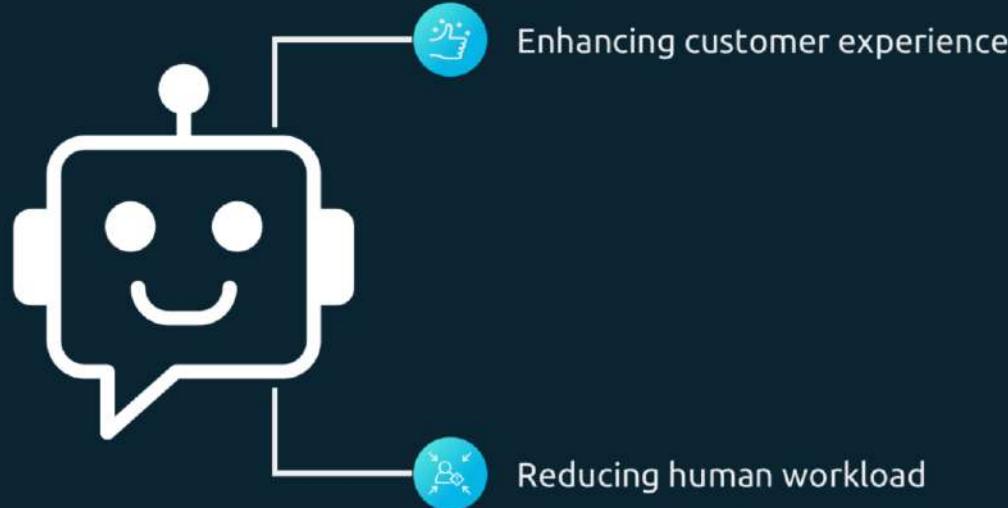
Fraud detection



Data analysis

Organizations use AI to automate tasks like customer service, fraud detection, and data analysis.

# Why Artificial Intelligence (AI) Matters



© Copyright KodeKloud

The demand for AI skills is growing because it enables businesses to gain insights and make faster, data-driven decisions. For example, AI-powered chatbots provide 24/7 customer support, enhancing customer experience while reducing human workload.

Understanding AI principles is crucial for any professional aiming to innovate using AWS AI services, such as Amazon SageMaker and Recognition.

# Why Artificial Intelligence (AI) Matters



Amazon SageMaker



Amazon Rekognition

© Copyright KodeKloud

Understanding AI principles is crucial for any professional aiming to innovate using AWS AI services, such as Amazon SageMaker and Rekognition.

# AI vs Rule-Based Systems



Rule-Based Systems



Machine Learning Systems

© Copyright KodeKloud

Rule-based systems operate on predefined logic, making them ideal for situations where the input and output need to be consistent and deterministic.

For example, a rule-based system could automatically approve loan applications for applicants with a credit score above 750. This is simple, transparent, and requires no AI intervention.

AI models, on the other hand, are probabilistic and adapt to data over time, making them suitable for more complex decisions where patterns evolve and uncertainty is involved.

It's important to evaluate whether a problem truly requires AI. If a business needs simple, repeatable outputs, a rule-based system may be a more cost-effective and reliable option.

## Artificial Intelligence (AI)

**AI is a branch of computer science focused on creating systems capable of performing tasks that typically require human intelligence.**

What is AI? AI is a branch of computer science focused on creating systems capable of performing tasks that typically require human intelligence.

# Artificial Intelligence (AI)



Visual  
perception



Speech  
recognition



Decision-  
making



Language  
translation

These include visual perception, speech recognition, decision-making, and language translation.

# Artificial Intelligence (AI)

01

Machine  
Learning (ML)

02

Deep Learning  
(DL)

AI encompasses various subfields, including Machine Learning (ML) and Deep Learning (DL).

# 01

Narrow AI (specific tasks)

# 02

General AI (broad capabilities)

AI can be categorized into Narrow AI (specific tasks) and General AI (broad capabilities).

# 01

## Narrow AI (specific tasks)



Alexa



Siri



Netflix



Amazon

Narrow AI: Focuses on performing specific tasks. Examples include Alexa, Siri, and recommendation engines on platforms like Netflix or Amazon.

# 02

## General AI (broad capabilities)



A theoretical form of AI with broad problem-solving abilities, much like a human

© Copyright KodeKloud

General AI: A theoretical form of AI with broad problem-solving abilities, much like a human.

# Artificial Intelligence (AI)

## Healthcare



Helps radiologists read X-rays faster and more accurately

## Retail



Makes product recommendations based on customers' previous shopping behavior, like recommending books on Amazon

© Copyright KodeKloud

## Examples:

AI in Healthcare: AI helps radiologists read X-rays faster and more accurately.

AI in Retail: AI makes product recommendations based on customers' previous shopping behavior, like recommending books on Amazon.

## Machine Learning (ML)

**ML involves using algorithms and statistical models to allow computers to perform tasks by learning from data rather than following explicit instructions.**

What is ML?: ML involves using algorithms and statistical models to allow computers to perform tasks by learning from data rather than following explicit instructions.

# Machine Learning (ML)



ML models improve automatically over time with more data.

# Machine Learning (ML)

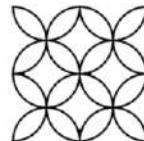


## Training Data

Large amounts of data are used to "train" the machine learning model.



KodeKloud



## Patterns and Predictions

The system identifies patterns and uses them to make predictions.

© Copyright KodeKloud

### How it Works:

**Training Data:** Large amounts of data are used to "train" the machine learning model.

# Machine Learning (ML)

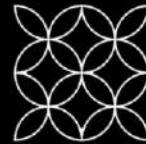


## Training Data

Large amounts of data are used to "train" the machine learning model.



KodeKloud



## Patterns and Predictions

The system identifies patterns and uses them to make predictions.

© Copyright KodeKloud

**Patterns & Predictions:** The system identifies patterns and uses them to make predictions, like recommending the next movie a user might want to watch on Netflix.

**Examples:**

**Spam Filtering:** Email services like Gmail use ML to automatically detect and filter out spam emails.

**Customer Service Bots:** Chatbots in customer service learn from interactions to provide better responses.

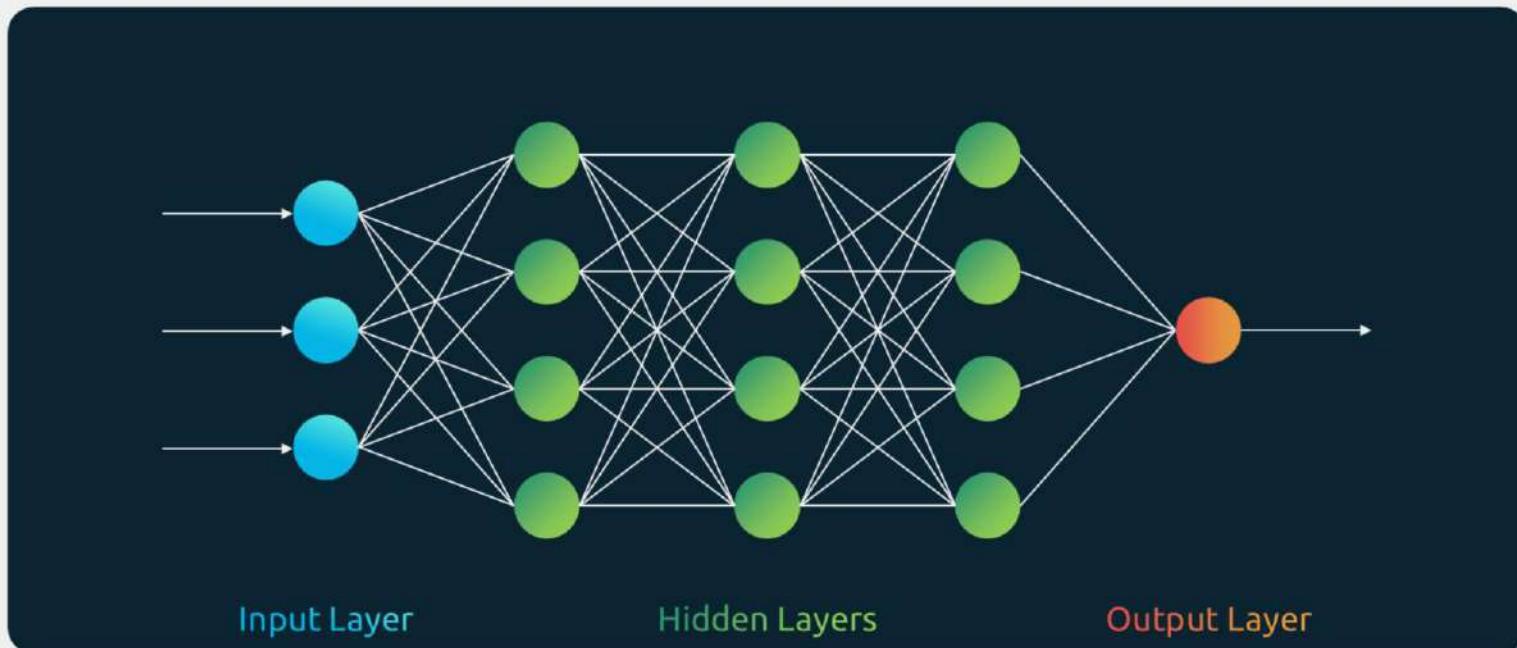
## Deep Learning

**Deep Learning is a subset of ML that uses multi-layered neural networks to model and solve more complex problems.**



What is Deep Learning?: Deep Learning is a subset of ML that uses multi-layered neural networks to model and solve more complex problems.

# Neural Networks



© Copyright KodeKloud

**Neural Networks:** These networks consist of layers of artificial neurons (nodes) that process input data through various layers, gradually extracting more abstract features.

**Applications:** DL is particularly powerful for tasks such as voice assistants (e.g., Alexa), facial recognition (e.g., Facebook's photo tagging), and language translation (e.g., Google Translate).

# Deep Learning

## Speech Recognition



Helps AI systems like Siri and Google Assistant understand and respond to voice commands

## Image Recognition



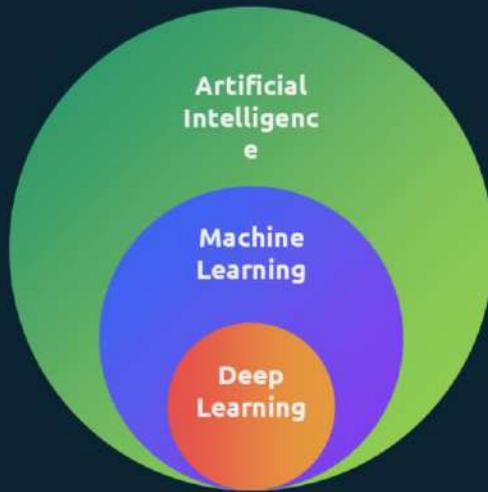
Used in systems like Google's image search, where AI identifies objects within an image

### Examples:

Speech Recognition: DL helps AI systems like Siri and Google Assistant understand and respond to voice commands.

Image Recognition: DL is used in systems like Google's image search, where AI identifies objects within an image.

# Comparing AI, ML, and Deep Learning



- A program that can sense, reason, act, and adapt
- Algorithms whose performance improve as they are exposed to more data over time
- Subset of ML in which multi-layered neural networks learn from vast amount of data

© Copyright KodeKloud

LLMs like OpenAI Codex are now assisting developers generating code snippets, right? This is basically not a human language, it's a coding language.



## **Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning – Similarities and Differences**

# Understanding AI, ML, and Deep Learning

01

## **Artificial Intelligence (AI)**

- Transforming industries
- Driving automation
- Enhancing productivity

02

## **Machine Learning (ML)**

- Learning from data patterns
- Making predictions

03

## **Deep Learning**

- Facial recognition
- Natural language processing

Why is it important to differentiate AI, ML, and Deep Learning?

Understanding the nuances between these terms helps organizations choose the right technology for their needs. AI (Artificial Intelligence), ML (Machine Learning), and Deep Learning are often used interchangeably, but they have distinct purposes and use cases.

By knowing these differences, businesses can make informed decisions about the tools they need to build efficient

solutions, be it a recommendation system, image recognition software, or automated customer service. The relevance of these technologies in modern businesses:

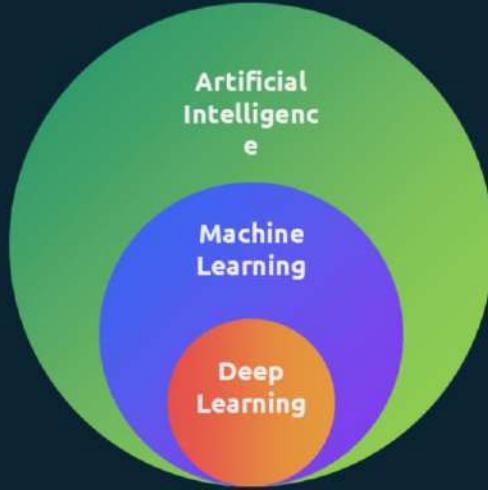
AI is transforming industries, driving automation, and enhancing productivity. Machine learning, a subset of AI, further enables businesses to learn from data patterns and make predictions. Deep Learning, a more advanced form of ML, powers cutting-edge applications such as facial recognition and natural language processing.

With the right application of these technologies, companies can gain a competitive edge, reduce costs, and provide better services.

How understanding the differences impacts decision-making and application:

Misunderstanding these concepts can lead to incorrect technology choices, increased costs, and ineffective solutions. For example, while AI can provide a general solution, ML is more suited for predictive analytics, and Deep Learning is ideal for complex pattern recognition.

# AI, ML, and Deep Learning – Key Similarities



- █ A program that can sense, reason, act, and adapt
- █ Algorithms whose performance improve as they are exposed to more data over time
- █ Subset of ML in which multi-layered neural networks learn from vast amount of data

© Copyright KodeKloud

All are interconnected fields under the AI umbrella:

AI is the overarching field, with ML as a subset that focuses on learning from data. Deep Learning is a further specialization within ML, emphasizing neural networks and complex data processing.

Data-driven nature:

Each of these technologies relies heavily on data to function effectively. The more data they process, the better they

perform. For example, both ML and Deep Learning models improve as they are trained with larger datasets.

Aim to automate and enhance decision-making:

AI, ML, and Deep Learning systems are designed to augment human capabilities by automating tasks, enhancing decision-making processes, and delivering intelligent solutions that are efficient and scalable.

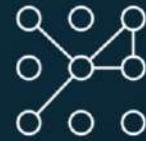
# AI, ML, and Deep Learning – Differences

Artificial Intelligence (AI)



General intelligence simulations

Machine Learning (ML)



Learning from data patterns

Deep Learning



Complex pattern recognition with neural networks

© Copyright KodeKloud

**AI: General intelligence simulations:**

AI encompasses a wide range of technologies and applications aimed at simulating human intelligence, from rule-based systems to self-learning algorithms.

**ML: Learning from data patterns:**

ML specifically involves developing models that learn from past data to predict or classify future events. Unlike general AI,

ML requires data to function effectively.

Deep Learning: Complex pattern recognition with neural networks:

Deep Learning goes beyond traditional ML by using neural networks that mimic the brain's structure, making it ideal for tasks that involve unstructured data like images and audio.

# Choosing the Right Technology

Application needs and complexity

Data availability and computational resources



© Copyright KodeKloud

Factors to consider when choosing between AI, ML, and Deep Learning:

It's essential to select the appropriate technology based on the problem at hand. For simple automation tasks, basic AI algorithms may be sufficient. For predictive analytics, ML models are most suitable, and for highly complex pattern recognition tasks like image processing, Deep Learning is the best choice.

Application needs and complexity:

Businesses need to assess the complexity of their application. Deep Learning models are typically more complex and require more resources, making them suitable for large-scale tasks that require high precision.

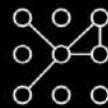
Data availability and computational resources:

ML and Deep Learning require substantial amounts of data and computational power. Organizations need to evaluate their resources and infrastructure before choosing to implement these technologies

# Choosing the Right Technology

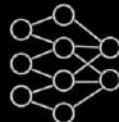
Application needs and complexity

Data availability and computational resources



## Machine Learning (ML)

Ideal for simpler tasks requiring predictive analysis



## Deep Learning

Suited for complex, large-scale tasks requiring high precision

© Copyright KodeKloud

Factors to consider when choosing between AI, ML, and Deep Learning:

It's essential to select the appropriate technology based on the problem at hand. For simple automation tasks, basic AI algorithms may be sufficient. For predictive analytics, ML models are most suitable, and for highly complex pattern recognition tasks like image processing, Deep Learning is the best choice.

Application needs and complexity:

Businesses need to assess the complexity of their application. Deep Learning models are typically more complex and require more resources, making them suitable for large-scale tasks that require high precision.

Data availability and computational resources:

ML and Deep Learning require substantial amounts of data and computational power. Organizations need to evaluate their resources and infrastructure before choosing to implement these technologies

# Choosing the Right Technology

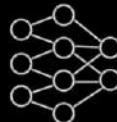
Application needs and complexity

Data availability and computational resources



## Machine Learning (ML)

Requires moderate data and resources



## Deep Learning

Needs substantial data and high computational power

© Copyright KodeKloud

Factors to consider when choosing between AI, ML, and Deep Learning:

It's essential to select the appropriate technology based on the problem at hand. For simple automation tasks, basic AI algorithms may be sufficient. For predictive analytics, ML models are most suitable, and for highly complex pattern recognition tasks like image processing, Deep Learning is the best choice.

Application needs and complexity:

Businesses need to assess the complexity of their application. Deep Learning models are typically more complex and require more resources, making them suitable for large-scale tasks that require high precision.

Data availability and computational resources:

ML and Deep Learning require substantial amounts of data and computational power. Organizations need to evaluate their resources and infrastructure before choosing to implement these technologies



# Types of Inferencing

# Inferencing in AI – Introduction



© Copyright KodeKloud

Now I realize that we keep saying models and this word model can be a bit misleading sometimes.

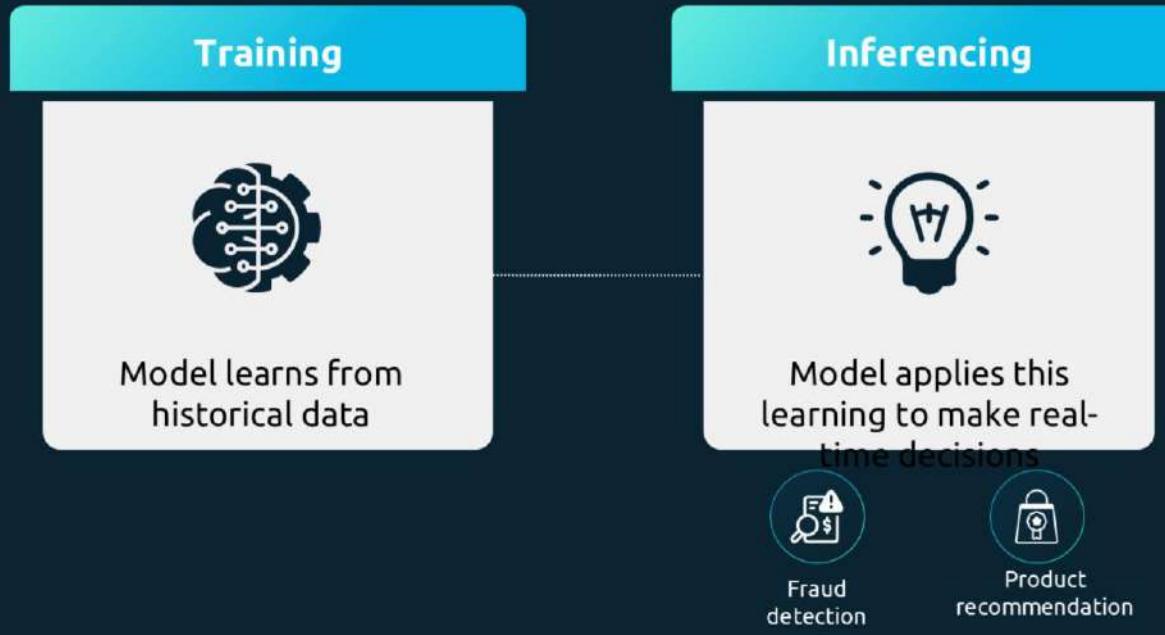
## Inferencing

**Inferencing is the process in which AI models make predictions or decisions using new data.**



**What is Inferencing?:** Inferencing is the use of trained AI models to generate predictions or classifications based on input data. After training, a model uses inferencing to provide outputs when exposed to new, unseen data.

# Inferencing in AI – Introduction



© Copyright KodeKloud

**Training vs Inferencing:** Training is when the model learns from historical data, while inferencing is when the model applies this learning to make real-time decisions.

**Example:** A spam filter that has been trained on millions of emails uses inferencing to determine if a new incoming email is spam.

Inferencing is the final application of a model's learning, helping businesses make real-time decisions such as fraud detection, recommendation systems, and customer service automation.

# Types of Inferencing

## Real-Time Inferencing

Processes data instantly as it arrives



Chatbots, fraud detection,  
autonomous driving systems



### Amazon SageMaker

Real-time endpoints for  
scalable and low-latency  
processing

## Batch Inferencing

Processes data in bulk at scheduled  
intervals



Sentiment analysis on social  
media posts collected over a day



### Amazon SageMaker

Transforms jobs for applying  
models to datasets in Amazon  
S3

© Copyright KodeKloud

## Real-Time Inferencing:

Real-time inferencing refers to the process of making predictions or classifications as data arrives. The model processes inputs instantly and returns outputs with minimal latency.

This is crucial for applications that require immediate responses, such as chatbots, fraud detection, autonomous driving systems, and recommendation engines. For example, when you interact with a virtual assistant like Amazon Alexa, the

system processes your voice input in real time to generate a relevant response.

In AWS, services such as Amazon SageMaker offer endpoints for deploying models that perform real-time inferencing. These endpoints are highly scalable and can handle varying workloads, ensuring that the system remains responsive even during high traffic periods.

#### Batch Inferencing:

In contrast, batch inferencing processes data in bulk. Rather than handling each data point as it arrives, the model processes multiple inputs simultaneously. This method is more efficient for scenarios where real-time responses are not necessary.

Examples include processing data from periodic customer surveys, performing large-scale sentiment analysis on social media posts collected over a day, or making predictions on data gathered over time (like stock market data analysis).

AWS services like SageMaker also support batch inferencing through batch transform jobs, which allow users to apply models to datasets stored in Amazon S3 efficiently.

# Real-Time Inferencing

- Involves making predictions instantly as data is received
- Common in applications such as voice assistants and self-driving cars



© Copyright KodeKloud

**What is Real-Time Inferencing?:** In this type, predictions or decisions are made immediately as input data is fed into the model. It's crucial for applications where quick responses are required.

**How it Works:** The AI model processes input data and delivers a prediction or classification in milliseconds.

**Applications:**

Voice Assistants: Systems like Alexa or Google Assistant process spoken language in real time and respond instantly.

Self-Driving Cars: AI systems process sensor data in real-time to make driving decisions like avoiding obstacles or following lanes.

Real-time inferencing is essential for applications requiring immediate feedback, like customer interactions, autonomous systems, or fraud prevention.

# Batch Inferencing

Involves processing large amounts of data at once, rather than in real time

Ideal for periodic analysis of data, such as overnight processing



KodeKloud



© Copyright KodeKloud

**What is Batch Inferencing?:** In batch inferencing, predictions or classifications are made on a large dataset at specific intervals, rather than as data is streamed in real time.

**How it Works:** Data is collected and processed in groups, often during off-peak hours or at scheduled times. The model processes the entire batch of data and generates results for all inputs simultaneously.

### Applications:

Customer Segmentation: A retailer might run a model at the end of each day to segment customers based on their shopping behavior.

Fraud Detection in Finance: Banks might run batch inferencing to analyze thousands of transactions overnight for fraud.

Batch inferencing is used when immediate responses aren't necessary, but large-scale data analysis is required. It's efficient for processing high volumes of data in one go.

# Batch Inferencing

Involves processing large amounts of data at once, rather than in real time

Ideal for periodic analysis of data, such as overnight processing



© Copyright KodeKloud

**What is Batch Inferencing?**: In batch inferencing, predictions or classifications are made on a large dataset at specific intervals, rather than as data is streamed in real time.

**How it Works:** Data is collected and processed in groups, often during off-peak hours or at scheduled times. The model processes the entire batch of data and generates results for all inputs simultaneously.

### Applications:

Customer Segmentation: A retailer might run a model at the end of each day to segment customers based on their shopping behavior.

Fraud Detection in Finance: Banks might run batch inferencing to analyze thousands of transactions overnight for fraud.

Batch inferencing is used when immediate responses aren't necessary, but large-scale data analysis is required. It's efficient for processing high volumes of data in one go.

# Batch Inferencing vs Real-Time Inferencing



Real-Time Inferencing

© Copyright KodeKloud

**Batch Inferencing:** In batch processing, the computing resources only run while a batch is being processed, allowing for cost savings. It is ideal for situations where results are not needed immediately, such as end-of-day financial reporting.

**Real-Time Inferencing:** For applications where immediate responses are needed (e.g., fraud detection in financial systems), real-time inferencing continuously runs to process incoming data.

### **Key Examples:**

**Batch:** Running daily customer segmentation at night to improve targeted marketing.

**Real-Time:** A real-time recommendation engine on an e-commerce website that suggests products as a customer browses.

# Batch Inferencing vs Real-Time Inferencing



Real-Time Inferencing

© Copyright KodeKloud

**Batch Inferencing:** In batch processing, the computing resources only run while a batch is being processed, allowing for cost savings. It is ideal for situations where results are not needed immediately, such as end-of-day financial reporting.

**Real-Time Inferencing:** For applications where immediate responses are needed (e.g., fraud detection in financial systems), real-time inferencing continuously runs to process incoming data.

### **Key Examples:**

**Batch:** Running daily customer segmentation at night to improve targeted marketing.

**Real-Time:** A real-time recommendation engine on an e-commerce website that suggests products as a customer browses.

# Batch Inferencing vs Real-Time Inferencing



Real-Time Inferencing



Batch Inferencing

© Copyright KodeKloud

**Batch Inferencing:** In batch processing, the computing resources only run while a batch is being processed, allowing for cost savings. It is ideal for situations where results are not needed immediately, such as end-of-day financial reporting.

**Real-Time Inferencing:** For applications where immediate responses are needed (e.g., fraud detection in financial systems), real-time inferencing continuously runs to process incoming data.

### **Key Examples:**

**Batch:** Running daily customer segmentation at night to improve targeted marketing.

**Real-Time:** A real-time recommendation engine on an e-commerce website that suggests products as a customer browses.

# Batch Inferencing vs Real-Time Inferencing



Real-Time Inferencing



Batch Inferencing

© Copyright KodeKloud

**Batch Inferencing:** In batch processing, the computing resources only run while a batch is being processed, allowing for cost savings. It is ideal for situations where results are not needed immediately, such as end-of-day financial reporting.

**Real-Time Inferencing:** For applications where immediate responses are needed (e.g., fraud detection in financial systems), real-time inferencing continuously runs to process incoming data.

### **Key Examples:**

**Batch:** Running daily customer segmentation at night to improve targeted marketing.

**Real-Time:** A real-time recommendation engine on an e-commerce website that suggests products as a customer browses.



# Artificial Data Types in AI Models

# Data Types in AI Models – Introduction

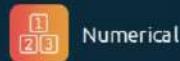


## Data Types in AI

**Data types refer to the various forms in which data can be represented and processed by AI models. These include numerical, categorical, and unstructured data.**

What are data types in AI? Data types refer to the various forms in which data can be represented and processed by AI models. These include numerical, categorical, and unstructured data. Each data type serves a different purpose in the AI model-building process.

# Data Types in AI Models – Introduction



Numerical



Categorical



Text



Images



AI Model

Vast amount  
of Data

Importance of data types in AI model training: Understanding and correctly handling data types is crucial for building accurate AI models. Each type requires different preprocessing and impacts how models learn from the data.

# Data Types in AI Models – Introduction



Amazon SageMaker

© Copyright KodeKloud

AWS provides various tools and services like Amazon SageMaker to manage these data types effectively for model training

# Numerical Data



KodeKloud



## Numerical Data

**Numerical data consists of quantitative values that can be measured and sorted in ascending or descending order.**



# Numerical Data



Integers



Floating-point  
numbers



Measurable  
quantities

These data types include integers, floating-point numbers, and any data that involves measurable quantities.

# Numerical Data

Time	Price (\$)	Sensor Data (Temp °C)
10:00	100	25.0
11:00	102	25.3
12:00	105	25.5



AI Model



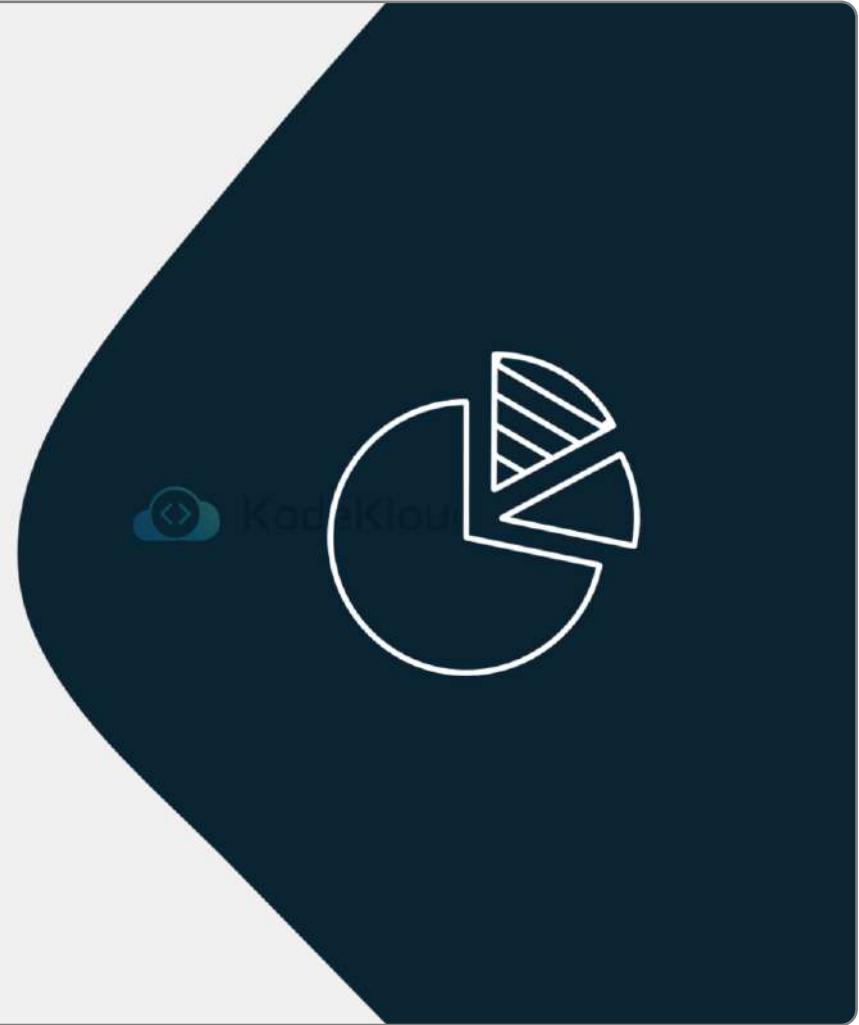
Financial forecasting



Sensor data analysis

Role of numerical data in AI models: Numerical data plays a significant role in machine learning models, particularly in regression analysis, neural networks, and statistical models. In AWS, services like SageMaker allow easy integration of numerical datasets into AI models. For example, numerical data is essential for predicting stock prices or analyzing sensor readings.

# Categorical Data



## Categorical Data

**Categorical data refers to information that can be divided into distinct categories or groups.**



KodeKloud

© Copyright KodeKloud

Definition of categorical data: Categorical data refers to information that can be divided into distinct categories or groups.

## Categorical Data



Gender



Product type



Geographical regions

Examples include gender (male, female), product type, and geographical regions.

# Categorical Data



© Copyright KodeKloud

Importance of encoding categorical data: AI models can't understand categorical data directly; it needs to be converted into a numerical format using techniques like one-hot encoding or label encoding. In AWS, Amazon SageMaker provides tools to preprocess categorical data for training models, making it easier for the machine to interpret this type of data.

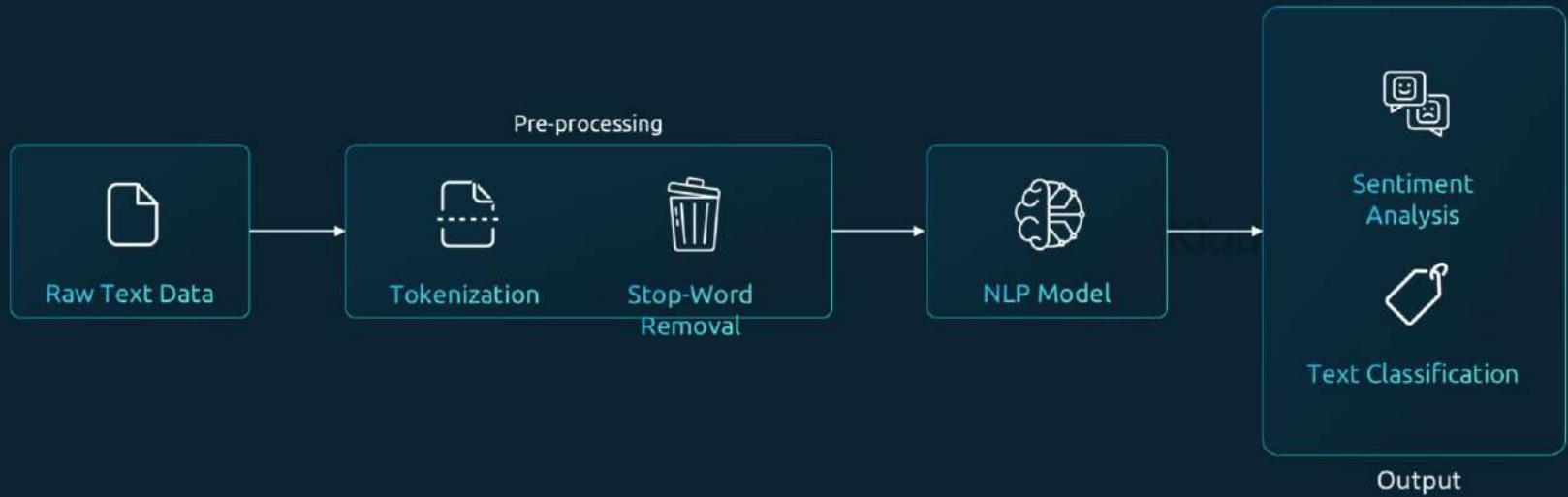
# Text Data and Natural Language Processing (NLP)



KodeKloud



# Text Data and Natural Language Processing (NLP)



© Copyright KodeKloud

**Text as unstructured data:** Text data is an example of unstructured data, which doesn't fit neatly into rows and columns. This type of data can be found in emails, social media posts, and articles.

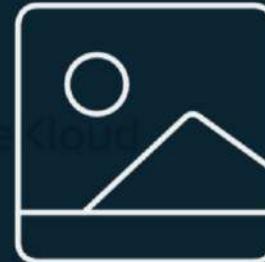
**How AI models process text data:** AI models process text data using Natural Language Processing (NLP). Services like Amazon Comprehend can extract meaning, detect sentiment, and understand language from unstructured text.

**Preprocessing steps:** Preprocessing steps such as tokenization, stop-word removal, and stemming/lemmatization are used to prepare text data for model training.

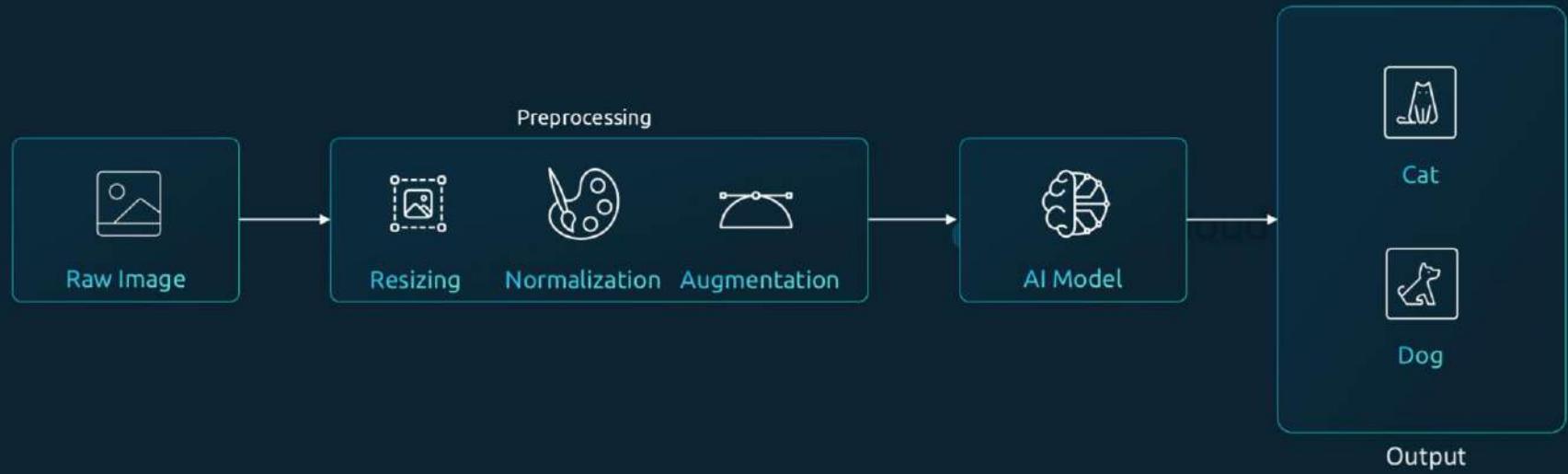
# Image Data



KodeKloud



# Image Data



© Copyright KodeKloud

Image data as unstructured data: Images are another form of unstructured data, consisting of pixels represented by numerical values. Unlike text or tabular data, image data requires special preprocessing techniques like resizing, normalization, and augmentation.

# Image Data



Image  
recognition



Object detection



Facial recognition

AI applications using image data: Image recognition, object detection, and facial recognition are common AI applications that rely on image data.

# Image Data

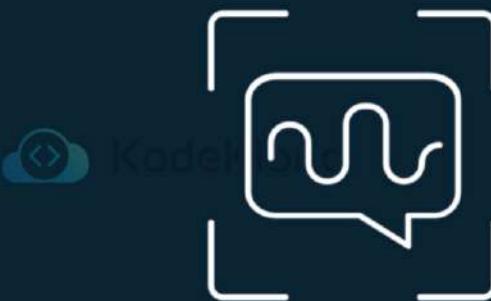


Amazon Rekognition

© Copyright KodeKloud

AWS services like Amazon Rekognition allow easy integration of image data into AI workflows for tasks such as identifying objects or detecting text in images.

# Audio Data and Speech Recognition



# Audio Data and Speech Recognition

Audio data is typically represented as time-series data, with amplitude values varying over time.



Speech  
recognition



Music analysis



Auditory  
applications

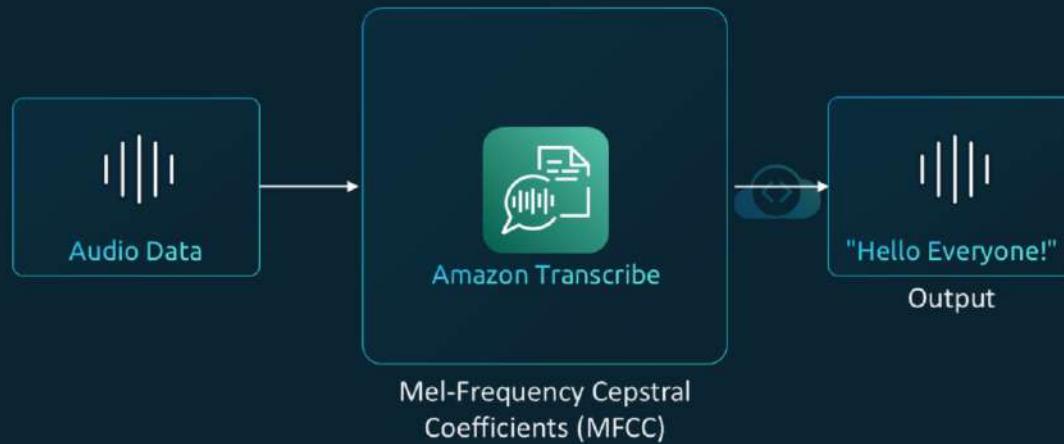
Audio data as time-series data: Audio data is typically represented as time-series data, with amplitude values varying over time. It's commonly used in speech recognition, music analysis, and other auditory applications.

# Audio Data and Speech Recognition



Speech recognition using AI models: Speech-to-text conversion is a critical AI application, relying on audio data.

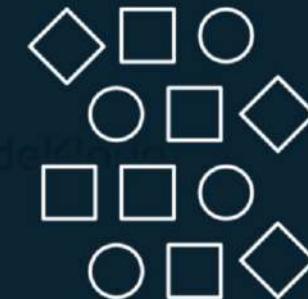
# Audio Data and Speech Recognition



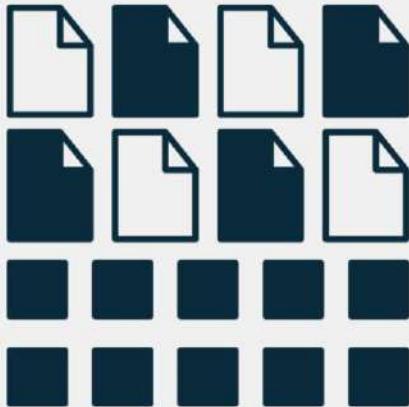
© Copyright KodeKloud

In AWS, Amazon Transcribe allows developers to convert audio files into text accurately. Preprocessing techniques such as feature extraction, including Mel-frequency cepstral coefficients (MFCC), are often used before feeding audio into models.

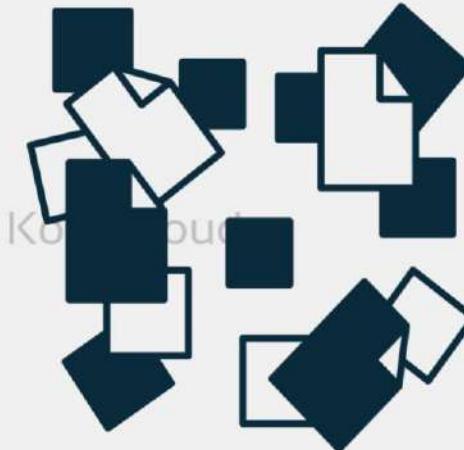
# Structured vs Unstructured Data



# Structured vs Unstructured Data



Structured Data



Unstructured Data

© Copyright KodeKloud

Differences between structured and unstructured data: Structured data is organized into clear formats like tables with rows and columns, while unstructured data lacks a predefined format (e.g., images, videos, free-text data).

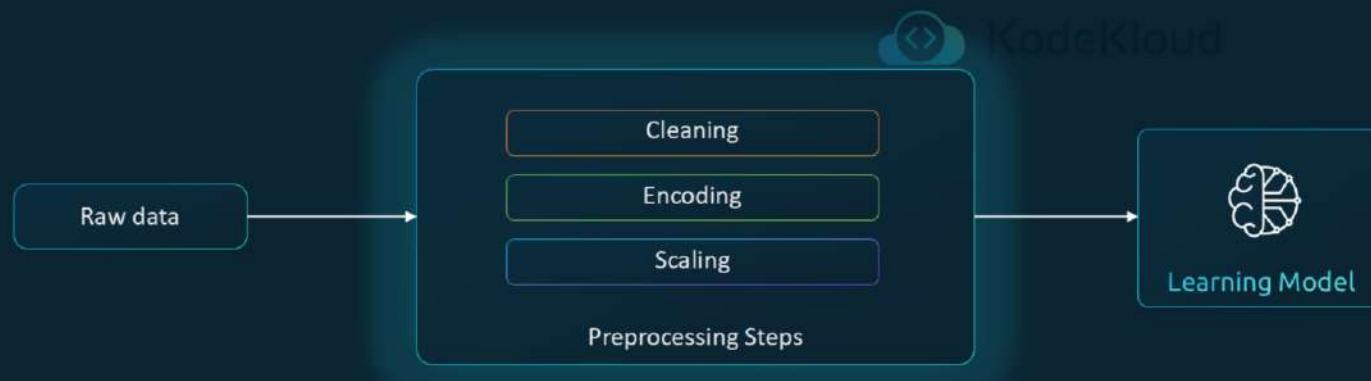
How AI models handle both data types: Structured data can be easily processed using traditional machine learning algorithms, while unstructured data, such as images and text, often requires advanced techniques like deep learning. AWS tools like SageMaker and Rekognition can process both types of data efficiently.

# Data Preprocessing for AI Models



# Data Preprocessing for AI Models

Poor-quality data can lead to inaccurate models, while well-preprocessed data ensure better performance.



© Copyright KodeKloud

**Importance of data preprocessing:** Data preprocessing is a crucial step in AI model development. Poor-quality data can lead to inaccurate models, while well-preprocessed data ensures better performance.

**Key preprocessing steps:** Key steps include data cleaning (removing noise or duplicates), normalization (scaling data to a consistent range), and transformation (encoding categorical data, feature scaling). AWS provides data preparation tools that help automate these tasks before the training phase.

# Labeled vs Unlabeled Data

## Labeled vs Unlabeled Data

**Labeled data contains both the input data and the corresponding output (target variable), while Unlabeled data has only the input without the expected output.**

# Labeled vs Unlabeled Data



© Copyright KodeKloud

Supervised vs. unsupervised learning: Supervised learning relies on labelled data to train models, making it ideal for tasks like classification and regression. Unsupervised learning uses unlabelled data, making it suitable for clustering and anomaly detection. AWS SageMaker supports both supervised and unsupervised learning models.

# Time-Series Data



## Time-Series Data

**Time-series data consists of data points collected or recorded at specific time intervals.**



What is time-series data? Time-series data consists of data points collected or recorded at specific time intervals. This type of data is used in stock market predictions, weather forecasting, and IoT device monitoring.

# Time-Series Data



ARIMA

Long Short-Term Memory  
networks (LSTM)

Prophet (a tool in AWS)

AI models that use time-series data: Time-series forecasting models like ARIMA, LSTM (Long Short-Term Memory networks), and Prophet (a tool in AWS) are used for predicting future trends based on past data. AWS provides tools for analyzing and forecasting time-series data effectively.

# Handling Imbalanced Data



## Imbalanced Data

**Imbalanced data refers to datasets where one class (in classification problems) has significantly more examples than the other(s). This can result in biased models that favor the dominant class.**

# Handling Imbalanced Data



Oversampling



Undersampling



ROC-AUC

Strategies to handle imbalanced datasets: Techniques such as oversampling, undersampling, and using performance metrics like ROC-AUC can help mitigate the effects of imbalanced data. AWS SageMaker includes built-in methods to address this issue during model training.

# Big Data and AI



Kode



# Big Data and AI

Big data refers to large and complex datasets that require advanced tools for processing.



Deeper  
insights



Train better  
models



Predict trends

How big data is used in AI models: Big data refers to large and complex datasets that require advanced tools for processing. AI models leverage big data to gain deeper insights, train better models, and predict trends.

# Big Data and AI



Numerical



Categorical



Text



Images



AI Model



How big data is used in AI models: Big data refers to large and complex datasets that require advanced tools for processing. AI models leverage big data to gain deeper insights, train better models, and predict trends.

# Big Data and AI



Amazon Elastic  
MapReduce (EMR)



Amazon SageMaker

Challenges of working with big data: Handling, storing, and processing large volumes of data can be resource-intensive. AWS offers tools like Amazon EMR (Elastic MapReduce) and SageMaker for handling big data effectively, allowing for the scaling of AI models to work with vast datasets.

# Handling Missing Data in AI Models



# Handling Missing Data in AI Models

Name	Age	Salary	Experience (Years)
Alice	25		3
Bob		50,000	5
Charlie	30		7
Diana		70,000	10

KodeKloud

Imputation technique used  
Mean and Median

Name	Age	Salary	Experience (Years)
Alice	25	60,000	3
Bob	27.5	50,000	5
Charlie	30	60,000	7
Diana	27.5	70,000	10

© Copyright KodeKloud

Why missing data is a problem: Missing data can reduce the quality and accuracy of AI models. If not handled properly, it may introduce bias or inaccuracies. AWS provides tools to automatically detect and handle missing data during the data preparation process.

Techniques to handle missing data: Some common techniques include imputing missing values with the mean, median, or mode of the data, or using algorithms that can handle missing data natively. SageMaker can incorporate these techniques automatically when preparing datasets for training models.

# Final Thoughts – Choosing the Right Data Type for AI Models

01

## Decision Trees

Structured Data  
(e.g., tabular data with numerical and categorical columns)

02

## CNNs

Unstructured Data  
(e.g., text, images)

03

## RNNs

Time-Series Data  
(e.g., stock prices over time)

Why is it important to differentiate AI, ML, and Deep Learning?

Understanding the nuances between these terms helps organizations choose the right technology for their needs. AI (Artificial Intelligence), ML (Machine Learning), and Deep Learning are often used interchangeably, but they have distinct purposes and use cases.

By knowing these differences, businesses can make informed decisions about the tools they need to build efficient

solutions, be it a recommendation system, image recognition software, or automated customer service. The relevance of these technologies in modern businesses:

AI is transforming industries, driving automation, and enhancing productivity. Machine learning, a subset of AI, further enables businesses to learn from data patterns and make predictions. Deep Learning, a more advanced form of ML, powers cutting-edge applications such as facial recognition and natural language processing.

With the right application of these technologies, companies can gain a competitive edge, reduce costs, and provide better services.

How understanding the differences impacts decision-making and application:

Misunderstanding these concepts can lead to incorrect technology choices, increased costs, and ineffective solutions. For example, while AI can provide a general solution, ML is more suited for predictive analytics, and Deep Learning is ideal for complex pattern recognition.



# **Supervised, Unsupervised, and Reinforcement Learning**

# Types of Machine Learning

## Supervised Learning

Model learns from labeled data



E.g.: Predicting stock market movements using historical data

## Unsupervised Learning

Model finds patterns in unlabeled data



E.g.: Identifying hidden customer segments based on purchasing behavior

## Reinforcement Learning

Model learns through trial and error by receiving rewards or penalties



E.g.: AlphaGo, the AI developed to play the board game Go, uses reinforcement learning

**Supervised Learning:** The model learns from labeled data (data with known outcomes). Example: Predicting housing prices based on features like square footage, location, and number of bedrooms.

**Unsupervised Learning:** The model finds patterns in unlabeled data. Example: Customer segmentation in marketing, where the system groups customers based on behavior without predefined labels.

**Reinforcement Learning:** The model learns through trial and error by receiving rewards or penalties. Example: AI playing chess, learning strategies through winning and losing games.

Examples:

Supervised: Predicting the stock market using historical data.

Unsupervised: Identifying hidden customer segments based on purchasing behavior.

Reinforcement: AlphaGo, the AI developed to play the board game Go, uses reinforcement learning.

# Supervised Learning



# Supervised Learning



© Copyright KodeKloud

**What is Supervised Learning?:** In supervised learning, models are trained using labeled datasets, meaning both the input and the desired output are known.

# Supervised Learning Example – Email Spam Detection



KodeKloud

© Copyright KodeKloud

## Email Spam Detection:

**Training Data:** Emails are labeled as either "spam" or "not spam."

**Model Training:** The model learns to identify patterns in spam emails, such as certain keywords or phrases that commonly appear in spam.

**Inferencing:** The model makes real-time inferences on new emails, assigning a probability that the email is spam.

**Challenges:** The model may need to be trained on millions of labeled emails to make accurate predictions, and it must continuously adapt as spammers find new techniques.

# Supervised Learning Example – Credit Scoring in Financial Institutions



© Copyright KodeKloud

## Credit Scoring:

**Training Data:** Historical data from past customers, including whether they defaulted or repaid loans.

**Model Training:** The model learns patterns that indicate high or low risk of default based on features like income, credit history, and employment.

**Inferencing:** When a new loan application is processed, the model predicts the likelihood of default, assigning a credit score.

Labeling data can be labor-intensive, as each application must be labeled as "good credit" or "bad credit," but it is essential to ensure the model provides accurate risk assessments.

# Unsupervised Learning



KodeKloud



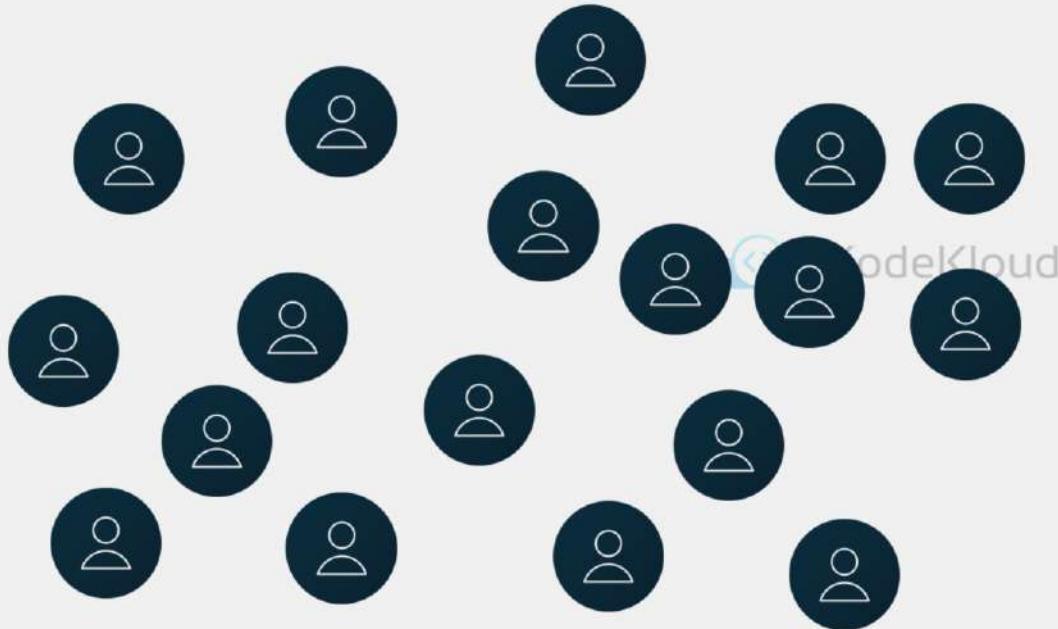
# Unsupervised Learning



© Copyright KodeKloud

What is Unsupervised Learning?: This method uses data that isn't labeled. It allows the model to find hidden patterns or groupings in the data.

# Unsupervised Learning Example – Customer Segmentation for Marketing



© Copyright KodeKloud

Training Data: Data from thousands of customers, including their purchase history, browsing behavior, and demographics.

Model Training: The model clusters customers into groups with similar purchasing habits without prior knowledge of the customer types.

Inferencing: The model helps the business design targeted marketing campaigns by identifying different customer segments (e.g., budget-conscious vs. premium buyers).

Unsupervised learning reduces the need for costly data labeling, enabling businesses to discover natural groupings in their customer base.

# Unsupervised Learning Example – Customer Segmentation for Marketing



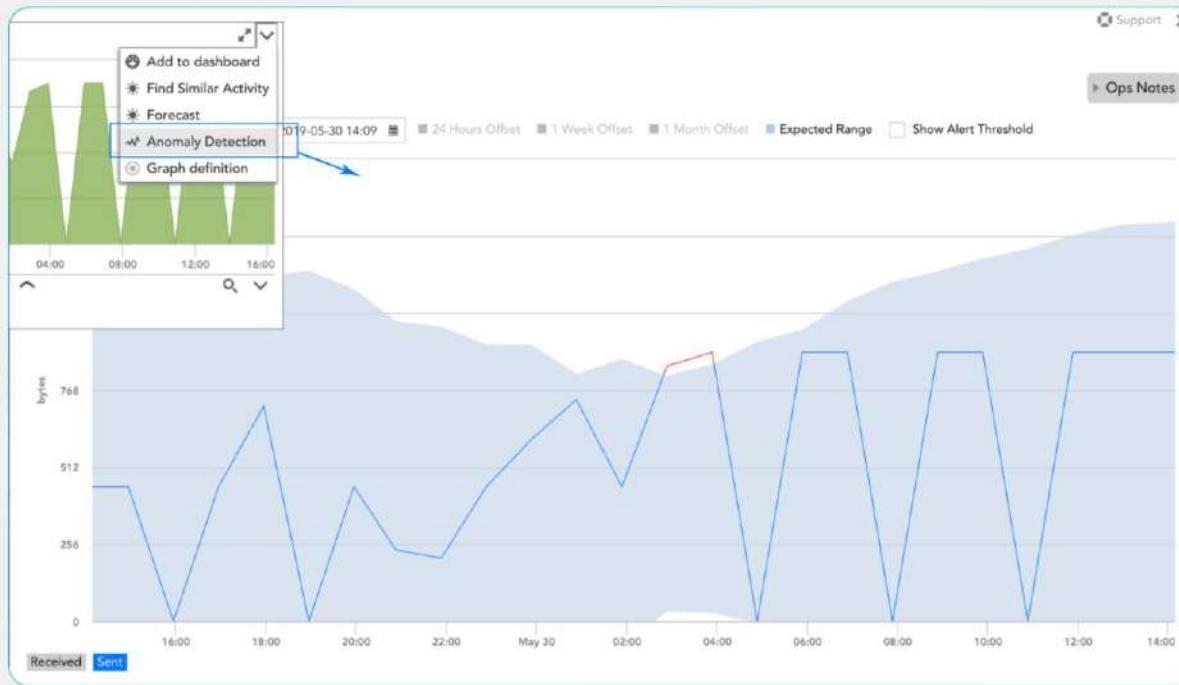
© Copyright KodeKloud

Training Data: Data from thousands of customers, including their purchase history, browsing behavior, and demographics.  
Model Training: The model clusters customers into groups with similar purchasing habits without prior knowledge of the customer types.

Inferencing: The model helps the business design targeted marketing campaigns by identifying different customer segments (e.g., budget-conscious vs. premium buyers).

Unsupervised learning reduces the need for costly data labeling, enabling businesses to discover natural groupings in their customer base.

# Unsupervised Learning Example – Anomaly Detection in Cybersecurity



© Copyright KodeKloud

Credits:

Training Data: Network traffic data without labels, including normal and potentially malicious behavior.

Model Training: The unsupervised learning algorithm clusters network traffic into normal and abnormal patterns.

Inferencing: The model identifies unusual network behavior (e.g., sudden spikes in traffic) and flags potential security threats.

While no labeling is required, the system must deal with false positives and continuously refine its ability to distinguish

between legitimate spikes and real threats.

# Reinforcement Learning



# Reinforcement Learning



© Copyright KodeKloud

Streaming Service Recommendations:

Agent: The recommendation engine.

Environment: The user's interaction with the platform.

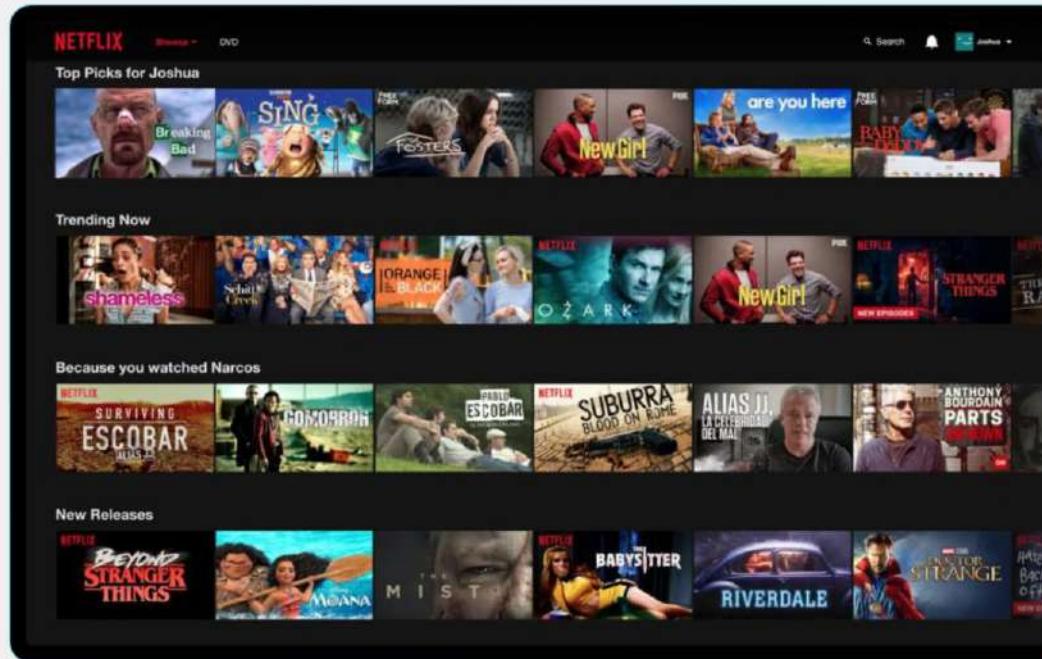
Action: The recommendation of a video, movie, or show.

Reward: Positive reward if the user watches the recommended content.

Inferencing: The system continuously learns from the user's behavior and updates its recommendations.

Reinforcement learning helps streaming platforms like Netflix or Amazon Prime optimize the user experience based on continuous feedback.

# Reinforcement Learning Example – Personalized Recommendations in Streaming Services



© Copyright KodeKloud

Credits:  
<https://www.netflix.com/in/>

Training Data: Network traffic data without labels, including normal and potentially malicious behavior.

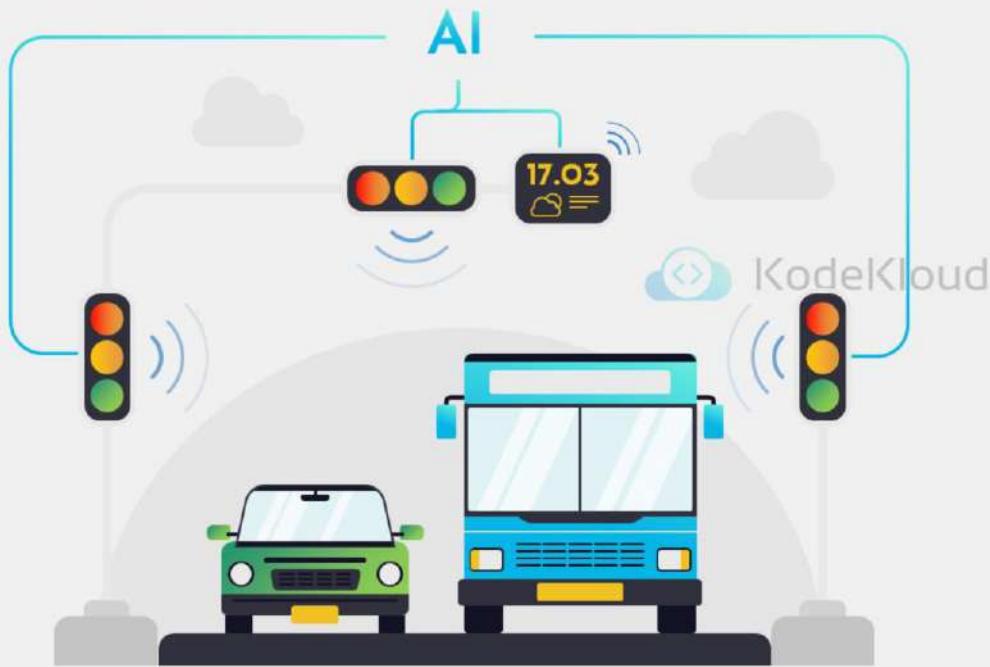
Model Training: The unsupervised learning algorithm clusters network traffic into normal and abnormal patterns.

Inferencing: The model identifies unusual network behavior (e.g., sudden spikes in traffic) and flags potential security threats.

While no labeling is required, the system must deal with false positives and continuously refine its ability to distinguish

between legitimate spikes and real threats.

## Reinforcement Learning Example – Traffic Light Optimization in Smart Cities



© Copyright KodeKloud

Traffic Light Optimization:

Agent: The AI managing the traffic lights.

Environment: Real-time traffic conditions.

Action: Changing the lights (red, yellow, green).

Reward: Reducing overall traffic congestion and improving vehicle flow.

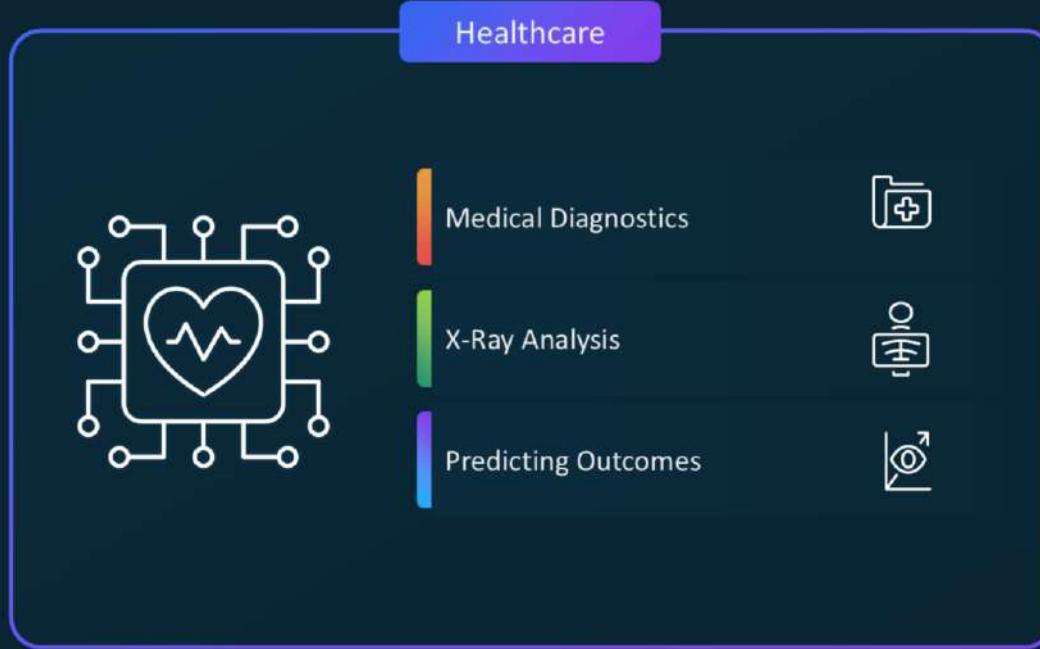
Inferencing: The system continuously adapts to changing traffic patterns, optimizing the sequence and duration of lights based on real-time data.

By using real-time data and continuous learning, the system improves traffic flow and reduces bottlenecks.



# Identifying Practice Use Cases in AI/ML

# Practical Applications of AI, ML, and DL



© Copyright KodeKloud

Healthcare: AI helps in medical diagnostics by reading X-rays, identifying diseases, and predicting patient outcomes.

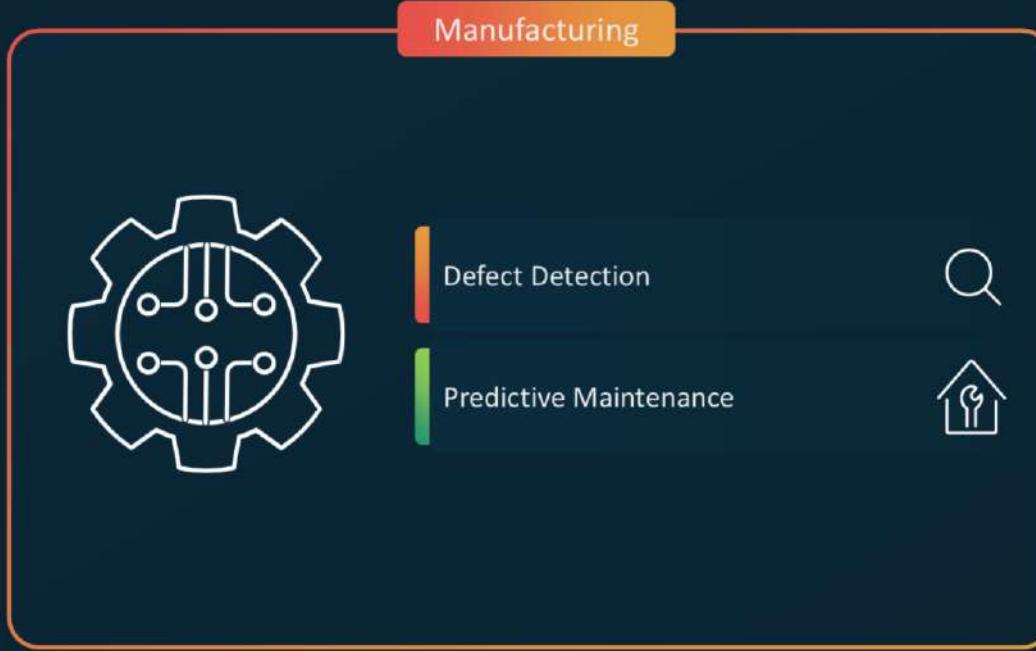
# Practical Applications of AI, ML, and DL



© Copyright KodeKloud

Finance: Machine learning is used to detect fraudulent transactions and offer personalized financial advice to customers.

# Practical Applications of AI, ML, and DL



© Copyright KodeKloud

Manufacturing: Deep learning algorithms use computer vision to detect defects on assembly lines and predict when equipment will fail (predictive maintenance).

# Practical Applications of AI, ML, and DL



Artificial Intelligence



Machine Learning



Deep Learning



Enhance efficiency



Reduce errors



Improve customer experiences

AI and its subsets offer businesses and industries powerful tools to enhance efficiency, reduce errors, and improve customer experiences.

# Identifying Practical Use Cases in AI

## Enhancing Operational Efficiency

01



Analyzes data continuously

02



Automates repetitive tasks

03



Boosts employee focus

AI's strength lies in its ability to operate continuously, analyzing large datasets and making predictions with high accuracy. Unlike human workers, AI doesn't experience fatigue, making it particularly useful for repetitive tasks.

Consider the role of AI in improving operational efficiency by automating mundane processes, allowing employees to focus on higher-level decision-making tasks.

# Identifying Practical Use Cases in AI



Customer Service  
Automation



Fraud  
Detection



Demand  
Forecasting

AI can simulate human-like intelligence in tasks that involve complex problem-solving, such as customer service automation, fraud detection, and demand forecasting

# AI in Customer Support



- Handles routine inquiries, freeing up human agents
- Responds to FAQs and troubleshooting steps
- Instant responses in e-commerce for order status, shipping, and returns

© Copyright KodeKloud

AI chatbots are increasingly used in customer service to handle routine inquiries, freeing up human agents for more complex tasks.

For example, AI can assist in responding to frequently asked questions or guiding users through troubleshooting steps without any human intervention.

A notable use case is e-commerce, where AI chatbots can provide instant responses about order status, shipping updates, or return policies.

# AI in Customer Support



Natural Language Processing  
(NLP)



Interacts in a human-like manner



Improves customer satisfaction



Reduces workload on human agents

By leveraging natural language processing (NLP), these AI models understand and respond to customer queries in a humanlike manner, enhancing customer satisfaction while reducing the workload on human agents.

# Practical Applications – Finance



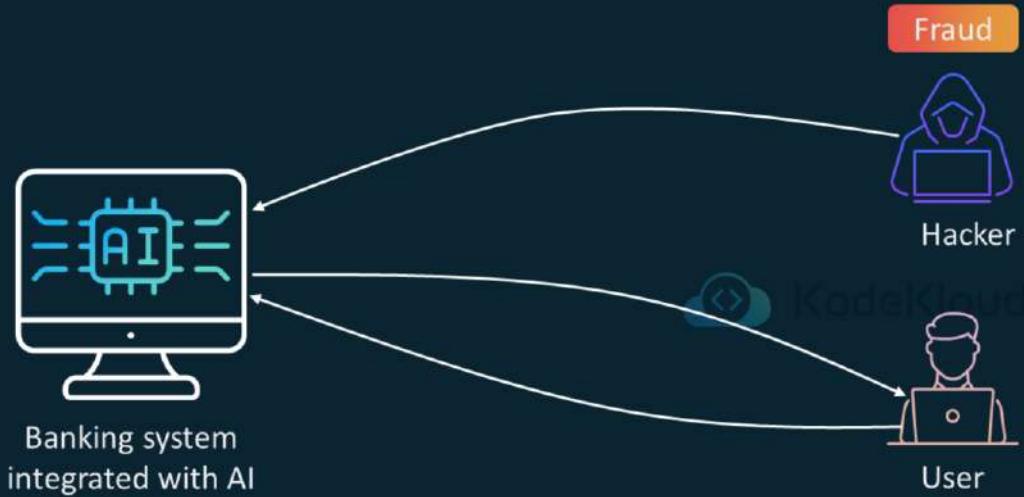
Detects fraudulent transactions



Offers personalized financial advice

Finance: Machine learning is used to detect fraudulent transactions and offer personalized financial advice to customers.

# AI in Fraud Detection



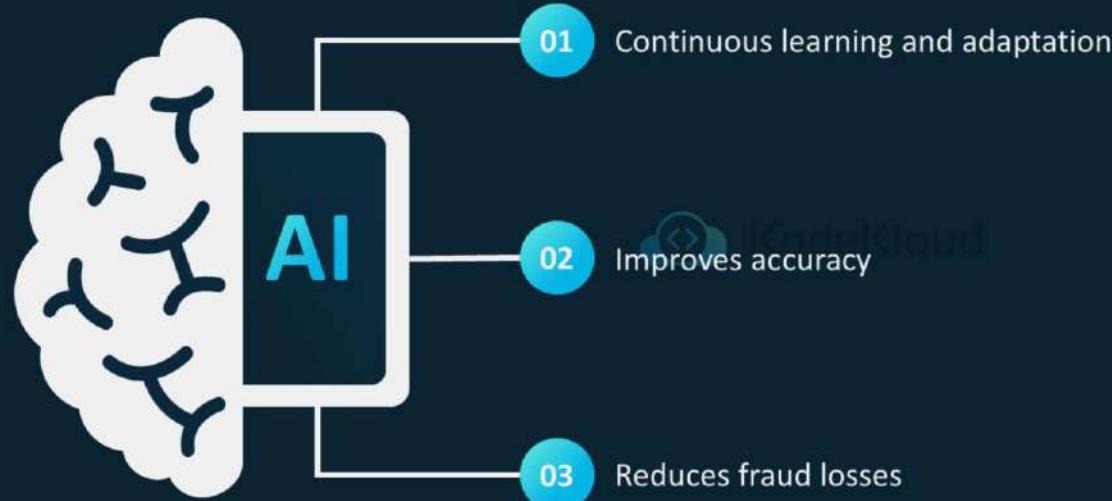
AI systems in banking monitor millions of transactions daily for suspicious activities.

© Copyright KodeKloud

AI excels in detecting anomalies in large sets of transactional data. By recognizing patterns and spotting deviations, AI systems can flag potentially fraudulent transactions.

In banking, AI systems monitor millions of transactions daily, identifying suspicious activity such as unusual spending patterns or account behaviors.

# AI in Fraud Detection



© Copyright KodeKloud

AI-based fraud detection models continuously learn and adapt, improving their accuracy over time, significantly reducing losses due to fraud.

For example, AI can detect when someone's credit card has been stolen by identifying purchases outside the usual behavior of the cardholder.

# AI in Predictive Maintenance



© Copyright KodeKloud

AI models can analyze data from sensors embedded in industrial machinery to predict when maintenance is required, preventing costly breakdowns.

By predicting failures before they occur, businesses can perform maintenance during scheduled downtimes, avoiding unexpected halts in production.

This is widely used in industries such as manufacturing and aviation, where unscheduled downtimes can be extremely costly.

For example, airlines use AI to predict when an engine part is likely to fail, allowing for proactive replacement before it causes any disruption.

# AI in Predictive Maintenance

01



Predicts failures

02



Maintenance  
during scheduled  
downtimes

03



Avoids unexpected  
halts

By predicting failures before they occur, businesses can perform maintenance during scheduled downtimes, avoiding unexpected halts in production.

# AI in Predictive Maintenance



Manufacturing



Aviation

© Copyright KodeKloud

This is widely used in industries such as manufacturing and aviation, where unscheduled downtimes can be extremely costly.

For example, airlines use AI to predict when an engine part is likely to fail, allowing for proactive replacement before it causes any disruption.

# Practical Applications – Healthcare



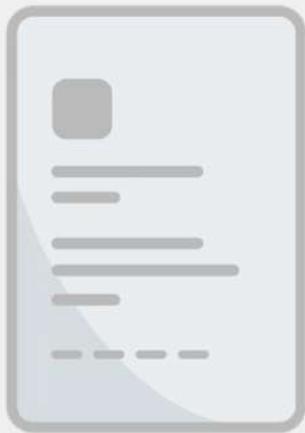
KodeKloud

AI helps in medical diagnostics by reading X-rays, identifying diseases, and predicting patient outcomes.

© Copyright KodeKloud

Healthcare: AI helps in medical diagnostics by reading X-rays, identifying diseases, and predicting patient outcomes.

# AI in Healthcare – Diagnosis and Treatment



1 | Medical images

2 | Patient records

3 | Genetic data

**Assist doctors in diagnosing diseases**

© Copyright KodeKloud

AI is transforming healthcare by analyzing medical images, patient records, and genetic data to assist doctors in diagnosing diseases.

# AI in Healthcare – Diagnosis and Treatment



Detects subtle patterns



Analyzes medical imaging



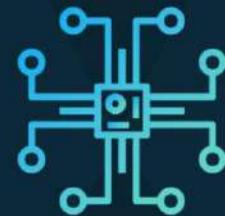
Develops personalized treatment plans

AI can detect subtle patterns in medical data that may go unnoticed by human eyes. For example, AI-powered diagnostic tools can analyze X-rays, MRIs, and CT scans to detect early signs of diseases such as cancer.

AI also helps in developing personalized treatment plans based on a patient's genetic profile, medical history, and lifestyle. For instance, AI can recommend a specific drug that has the highest probability of success for a patient based on prior treatment outcomes from similar cases.

# AI in Demand Forecasting

- Historical sales data
- Market trends
- Consumer behavior patterns



AI can predict future product demand by analyzing historical sales data, market trends, and consumer behavior patterns.

# AI in Demand Forecasting



- | Forecast demand 👁️
- | Reduce waste 🗑️
- | Prevent lost sales 🛒

© Copyright KodeKloud

Retail companies use AI to forecast demand and ensure that the right amount of stock is available at the right time. This reduces waste due to overstocking and prevents lost sales due to understocking.

For example, a grocery chain might use AI to predict how much fresh produce to stock based on seasonal trends, holidays, and local events.

AI models adapt in real-time, making it possible to respond quickly to changes in demand due to unexpected factors like weather conditions or economic shifts.

# AI in Autonomous Vehicles



Sensors



Cameras



Radars



KodeKloud



© Copyright KodeKloud

Autonomous vehicles rely on AI to process data from sensors, cameras, and radars to understand their surroundings and make decisions in real time.

# AI in Autonomous Vehicles



© Copyright KodeKloud

AI in self-driving cars identifies obstacles, traffic signals, pedestrians, and other vehicles, making real-time driving decisions that mimic human judgment.

For example, Tesla's Autopilot uses AI to navigate highways, avoid collisions, and even park the car autonomously. The transportation industry is investing heavily in AI to bring fully autonomous vehicles to the market, which could drastically reduce accidents caused by human error.

# AI in Agriculture

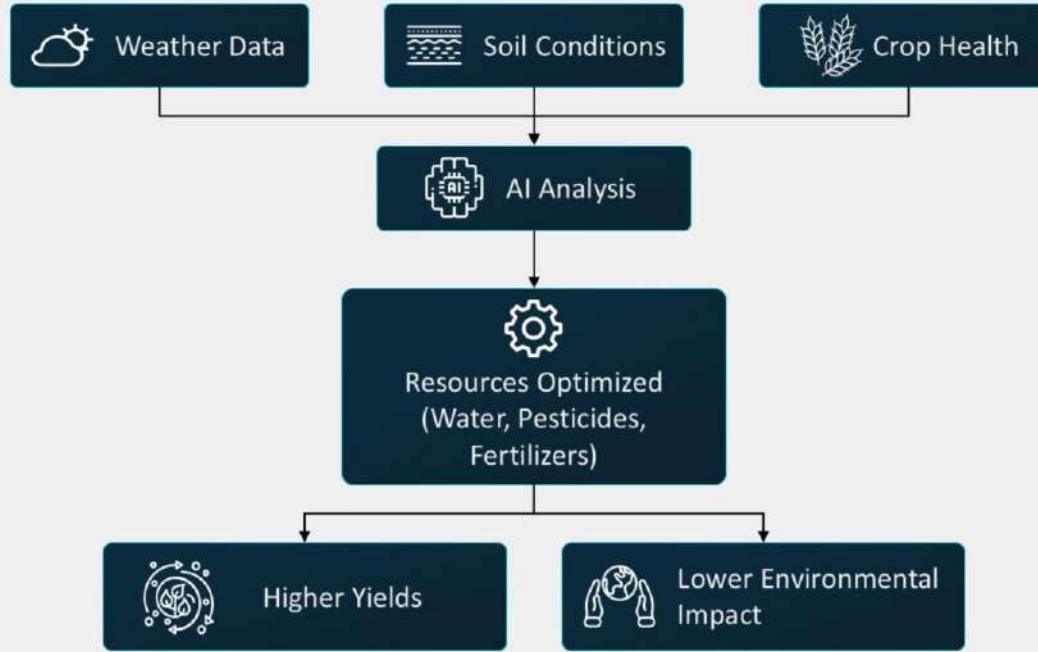


AI assists farmers in making decisions about planting, watering, fertilizing, and harvesting crops.

© Copyright KodeKloud

AI helps farmers make decisions about when and where to plant, how much water and fertilizer to use, and when to harvest crops.

# AI in Agriculture



© Copyright KodeKloud

By analyzing weather data, soil conditions, and crop health, AI can optimize the use of resources like water, pesticides, and fertilizers, leading to higher yields and lower environmental impact.

For example, AI-powered drones can survey fields, collecting data that helps farmers make precise decisions about crop management.

AI also helps in pest detection and disease prevention, enabling farmers to take early action before issues become widespread.

# Practical Applications – Manufacturing



Deep learning algorithms use computer vision to detect assembly line defects and predict equipment failures.

© Copyright KodeKloud

Manufacturing: Deep learning algorithms use computer vision to detect defects on assembly lines and predict when equipment will fail (predictive maintenance).

# AI in Robotics – Automation of Physical Tasks



## AI-Enabled Precision and Adaptability

Tasks such as factory assembly, warehouse sorting, and package delivery



## 24/7 Operation

Increased production rates while maintaining quality control



## Real-Time Adjustments

AI-driven robots optimize performance by learning efficient routes

AI enables robots to perform tasks that require precision and adaptability, such as assembly lines in factories, warehouse sorting, and even delivering packages.

Robotics in manufacturing can operate 24/7, significantly increasing production rates while maintaining quality control. AI-powered robots can adapt to changes in their environment, making real-time adjustments to optimize performance. For instance, in a warehouse, AI-driven robots can learn the most efficient routes for picking and delivering items. The use of AI in robotics extends beyond industrial settings. Autonomous robots are also being used in agriculture,

healthcare (for surgeries), and hospitality (as service robots).

# AI in Daily Life



**AI-driven services make life more convenient and efficient.**

© Copyright KodeKloud

AI has become deeply integrated into everyday life, often in ways that enhance convenience and user experience.

# Virtual Assistants



Google Home

© Copyright KodeKloud

Virtual Assistants: Devices like Amazon Alexa or Google Home use AI to understand voice commands and perform tasks like setting reminders or controlling smart home devices.

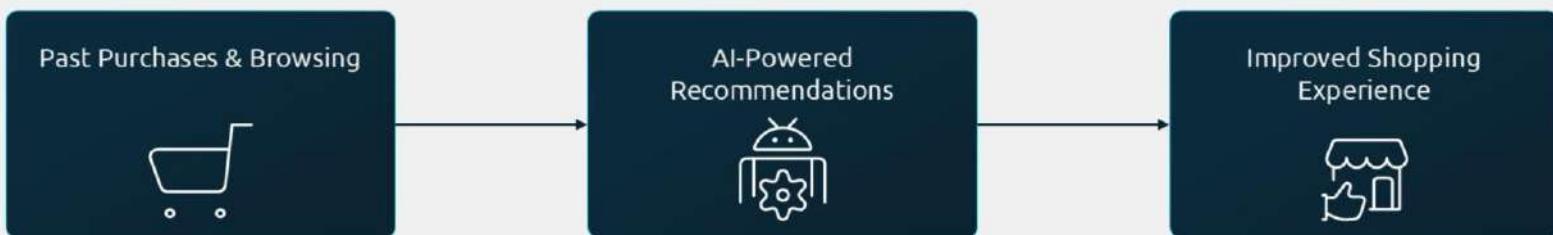
## Streaming Services



© Copyright KodeKloud

Streaming Services: Platforms like Netflix and YouTube use AI to recommend content based on your viewing habits.

# Online Shopping

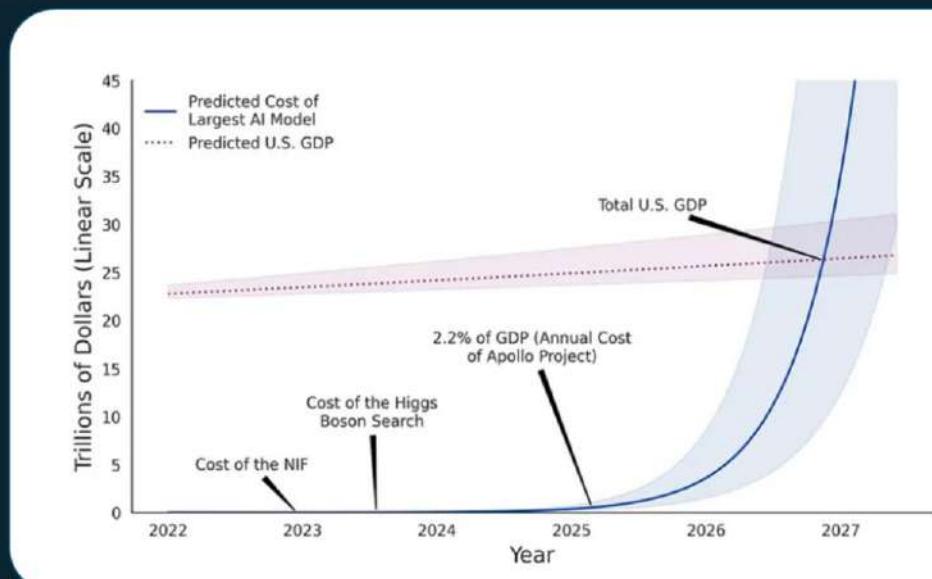


© Copyright KodeKloud

Online Shopping: Retailers use AI to recommend products based on your past purchases and browsing behavior, making shopping more personalized.

# Limitations of AI – High Costs

- High costs of AI implementation 
- Regular retraining needed 
- Businesses must assess benefits vs costs 
- Ongoing expenses: Storage, cloud, staff 



Implementing AI solutions can be costly due to the significant computational power required to train models, especially in machine learning and deep learning.

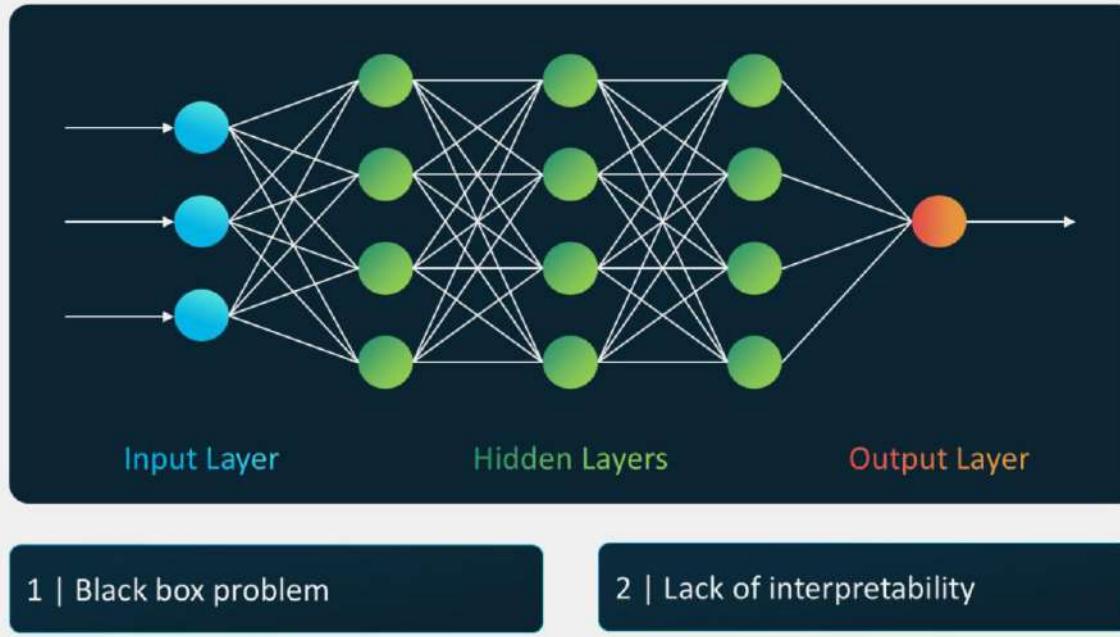
AI models need to be retrained regularly to remain effective, which requires additional resources and time.

Businesses need to assess whether the financial benefits of implementing AI outweigh the costs. For example, if the cost to train a fraud detection model is higher than the savings it generates, the project may not be viable.

Companies must also consider ongoing operational costs, such as data storage, cloud services, and expert personnel to

maintain AI systems.

# Limitations of AI – Lack of Interpretability



© Copyright KodeKloud

One of the major limitations of AI, particularly in deep learning, is that it can be difficult to understand how the model arrives at its conclusions. This is known as the “black box” problem.

For example, a neural network may correctly predict that a customer will default on a loan, but it may be impossible to explain the exact reasoning behind the prediction.

This lack of interpretability poses a problem in industries like finance and healthcare, where decisions need to be transparent for regulatory reasons.

In some cases, simpler, rule-based systems may be preferred because they offer clearer explanations for their decisions, even if their performance is lower than more complex AI models.



# ML Development Lifecycle and the ML Pipeline

# ML Development Lifecycle – Introduction



Machine learning models are dynamic and require continuous updates and retraining.

© Copyright KodeKloud

The ML lifecycle involves several interconnected stages, starting with defining a business goal and concluding with monitoring a deployed model.

Key phases include data collection, model training, deployment, and ongoing monitoring.

The ML pipeline is dynamic, often requiring iteration as new data becomes available or model performance degrades over time.

AWS provides multiple services to support each stage of this lifecycle

# ML Development Lifecycle – Introduction



Provides multiple services to support each stage of this lifecycle

© Copyright KodeKloud

The ML lifecycle involves several interconnected stages, starting with defining a business goal and concluding with monitoring a deployed model.

Key phases include data collection, model training, deployment, and ongoing monitoring.

The ML pipeline is dynamic, often requiring iteration as new data becomes available or model performance degrades over time.

AWS provides multiple services to support each stage of this lifecycle

# Business Goal Identification



1 | Increase customer retention

2 | Boost revenue by 15%

3 | Reduce operational costs by 10%

Every ML project must begin with a well-defined business goal to solve a specific problem.

# Business Goal Identification



Business  
Goal



Success is measured against  
business objectives

# Business Goal Identification



Business  
Goal



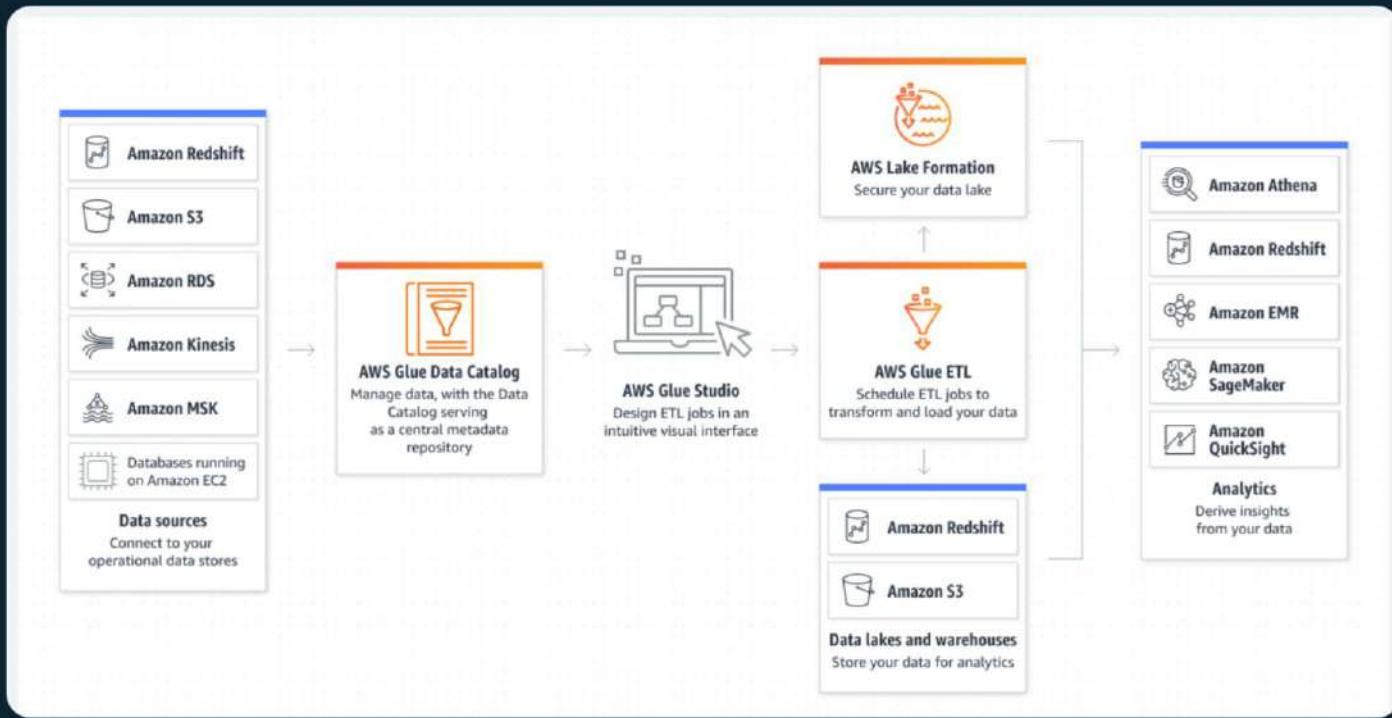
Aligns stakeholders with clear  
goals

Stakeholder alignment is crucial for setting measurable business objectives that justify moving the project forward.

# 01

## Data Collection and Preparation

# Data Collection and Preparation



© Copyright KodeKloud

Source: [https://d1.awsstatic.com/products/aws-glue/product-page-diagram\\_AWS-Glue\\_Elixir%20\(2\).522ef785088de982530b9fdde4c8be146562fa0f.png](https://d1.awsstatic.com/products/aws-glue/product-page-diagram_AWS-Glue_Elixir%20(2).522ef785088de982530b9fdde4c8be146562fa0f.png)

Data collection involves identifying relevant data sources and determining whether to process the data in real time (streaming) or batch mode.

The process includes ETL (Extract, Transform, Load) jobs to centralize data storage.

Source: [https://d1.awsstatic.com/products/aws-glue/product-page-diagram\\_AWS-Glue\\_Elixir%20\(2\).522ef785088de982530b9fdde4c8be146562fa0f.png](https://d1.awsstatic.com/products/aws-glue/product-page-diagram_AWS-Glue_Elixir%20(2).522ef785088de982530b9fdde4c8be146562fa0f.png)

# Data Collection and Preparation



Amazon S3



AWS Glue



Amazon Kinesis



AWS Lambda



Amazon Redshift



Amazon Elastic  
MapReduce (EMR)

© Copyright KodeKloud

## AWS Services:

**Amazon S3:** Used as the central data storage service where all collected and processed data is stored.

**AWS Glue:** A fully managed ETL service that allows you to perform data transformation and loading into Amazon S3 or other destinations. It's effective for batch processing.

**Amazon Kinesis:** For real-time or streaming data processing. You can use Kinesis Data Streams for ingesting streaming data

and Kinesis Data Firehose for loading the data into S3.

AWS Lambda: For serverless processing of incoming data, often used in combination with Kinesis for real-time processing.

Amazon Redshift: If you need to process large volumes of structured data and perform complex queries, Redshift can be used as a data warehouse.

Amazon EMR (Elastic MapReduce): Useful for big data processing using frameworks like Apache Hadoop and Apache Spark.

# Data Preprocessing and Feature Engineering



- | Cleaning and normalizing data 🧹
- | Handling missing values 📝
- | Transforming data 🔄

Data preprocessing involves cleaning and normalizing data, handling missing values, and transforming data to make it usable for model training.

# Data Preprocessing and Feature Engineering



**Data preprocessing and feature engineering optimize the data for training.**

# Data Preprocessing and Feature Engineering



AWS Glue



Amazon SageMaker

© Copyright KodeKloud

## AWS Services:

**AWS Glue:** This fully managed ETL service can be used for data cleaning, normalization, and transformation. Glue jobs can automate preprocessing tasks and handle missing values.

**Amazon SageMaker Processing:** This service allows you to run data preprocessing and feature engineering scripts in a

managed environment, using frameworks like Python with Pandas, Apache Spark, or Scikit-learn. It's particularly effective for handling large datasets in a scalable manner.

# Data Augmentation in AI Models

**Artificially increases dataset size through transformations**



Original image



Flipped image



Rotated image



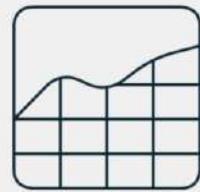
Cropped image



Increased  
data size

What is data augmentation? Data augmentation is the process of artificially increasing the size of a dataset by applying transformations such as flipping, rotating, or cropping images. This helps improve the diversity of the training data.

# How Data Augmentation Improves Model Performance



Better Generalization



Improved Accuracy

How it improves model performance: Data augmentation is particularly useful in image recognition models, where more varied data can lead to better generalization and improved accuracy.

# Data Augmentation in AI Models



Amazon SageMaker

Supports augmentation techniques to **enhance training datasets**

# Splitting Data for Training, Validation, and Testing



Training



Used to teach the model



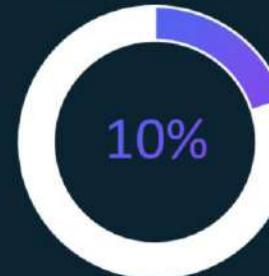
Validation



Used for fine-tuning



Testing



Measures final performance

© Copyright KodeKloud

Data should be split into three sets: training, validation, and testing, with a common split being 80% for training, 10% for validation, and 10% for testing.

Training data is used to teach the model, validation data is used for tuning, and testing data evaluates final model performance before deployment.

# Splitting Data for Training, Validation, and Testing



**Amazon SageMaker**

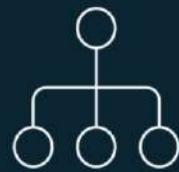
Data handling tools **streamline** and **automate** the data splitting process.

SageMaker's data handling tools make it easy to automate data splitting processes.

# 02

## Training the Model

# Training the Model



Automated Resource Management



SageMaker



Algorithm and Framework Support

During training, the model learns by adjusting weights based on errors between predicted and actual outcomes. SageMaker makes it easy to create training jobs using pre-defined algorithms or custom-built models. The training process is iterative, optimizing parameters to minimize prediction error. It's important to run multiple experiments with different algorithms and hyperparameters to achieve the best results.

# Hyperparameter Tuning



© Copyright KodeKloud

Hyperparameters control aspects of the model that affect performance, such as learning rate or neural network architecture.

# Hyperparameter Tuning

01



Enhances  
efficiency

02



Reduces  
error

Tuning is essential to find the best set of hyperparameters that minimize error and maximize performance.

# Hyperparameter Tuning



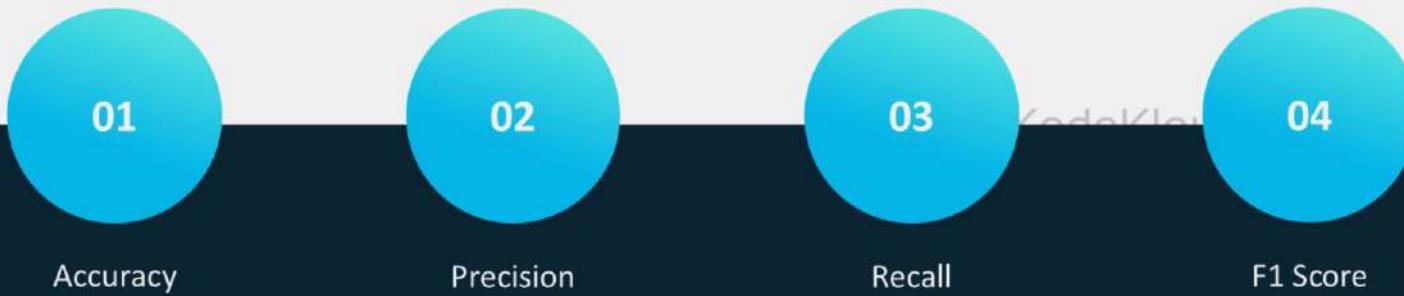
Amazon SageMaker  
Automatic Model Tuning (AMT)

Auto-optimize models by running multiple training jobs  
with different **hyperparameter** configurations.

© Copyright KodeKloud

SageMaker's Automatic Model Tuning (AMT) allows users to optimize models automatically by running multiple training jobs with different hyperparameter configurations.

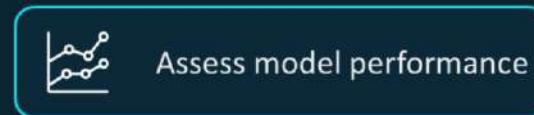
# Evaluating Model Performance



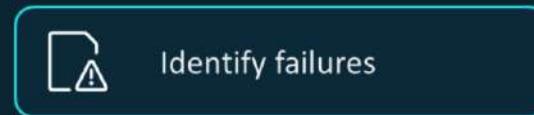
Once trained, models must be evaluated using specific performance metrics such as accuracy, precision, recall, and F1 score.

# Evaluating Model Performance

		Actual Value
		Positive
Predicated Value	Positive	True Positive
	Negative	False Positive
Predicated Value	Negative	False Negative
	Negative	True Negative



Assess model performance



Identify failures

Confusion matrices and other visual tools help in assessing model performance and identifying areas where it may be failing.

# Evaluating Model Performance



**Amazon SageMaker**

Built-in tools **evaluate models** and **track performance** against business metrics.

SageMaker provides built-in tools to evaluate models and track their performance against business metrics.

# 03

## Model Deployment

# Model Deployment Options



## Real-Time

Responds instantly to input



## Batch

Processes large amounts of data periodically

After training and evaluation, models are deployed for inference, which can be real-time (responding instantly to input) or batch (processing large amounts of data periodically).

# Model Deployment Options



Amazon SageMaker



AWS Lambda



Amazon Elastic  
MapReduce (EMR)

© Copyright KodeKloud

AWS offers multiple deployment options, including AWS Lambda for serverless inference, Amazon ECR (for containerized models), and Amazon SageMaker for managed endpoints, and AWS Batch for large-scale batch processing. The choice between batch and real-time depends on business requirements for speed, cost, and scalability.

# 04

## Monitoring Deployed Models

# Monitoring Deployed Models



Continuous monitoring is essential to ensure models perform as expected.

© Copyright KodeKloud

After deployment, continuous monitoring is essential to ensure models perform as expected. Models can suffer from data drift or concept drift, which can degrade performance over time.

# Monitoring Deployed Models



Amazon SageMaker



Amazon CloudWatch

© Copyright KodeKloud

Amazon SageMaker Model Monitor continuously checks for deviations in model behavior and alerts for necessary retraining. Automated retraining cycles can be initiated based on predefined triggers to ensure models remain accurate and up-to-date.

# Integrating the ML Pipeline for Success



© Copyright KodeKloud

The ML development lifecycle is a continuous, iterative process requiring attention to data, model performance, and business goals.

AWS services like SageMaker, Glue, and DataBrew offer powerful tools to streamline the entire process, from data collection to model deployment and monitoring.

By leveraging these tools, organizations can build scalable, reliable ML solutions that drive measurable business outcomes.



# MLOps Concepts From Design to Metrics – Introduction

# MLOps Concepts – Introduction



© Copyright KodeKloud

MLOps, short for Machine Learning Operations, integrates the principles of DevOps with machine learning, enabling seamless collaboration between data scientists and operations teams.

# MLOps Concepts – Introduction



KodeKloud



Automates ML deployment,  
monitoring, and updates

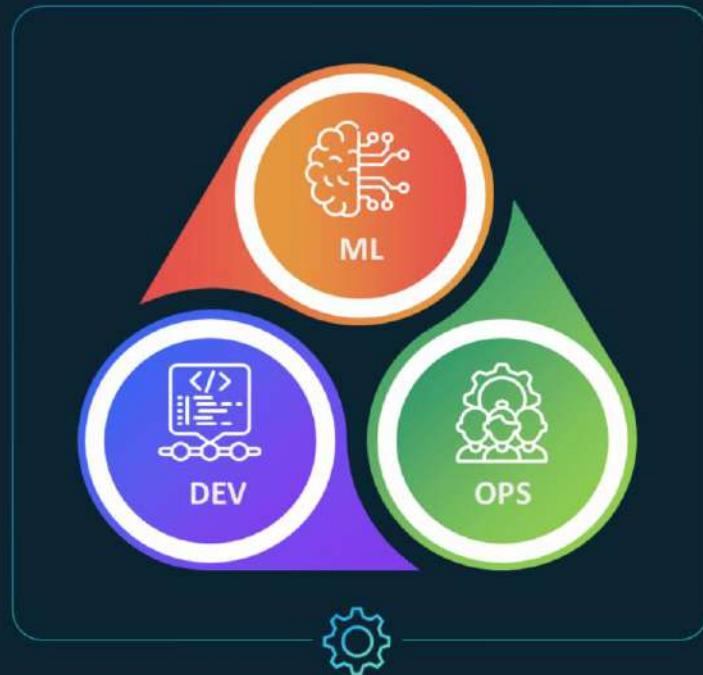


Uses CI/CD for continuous  
model delivery

© Copyright KodeKloud

It focuses on automating and optimizing all the processes involved in deploying, monitoring, and updating ML models. The core idea is to manage machine learning like software, ensuring continuous delivery and continuous integration (CI/CD) of models into production environments.

# Designing an MLOps Pipeline



© Copyright KodeKloud

MLOps pipelines automate every step, from data preparation to model training and deployment.

# Designing an MLOps Pipeline



© Copyright KodeKloud



Amazon  
SageMaker



Apache  
Airflow

Orchestrates complex workflows

These pipelines are built using tools like Amazon SageMaker Pipelines or Apache Airflow, which allow orchestration of complex workflows.

# Designing an MLOps Pipeline

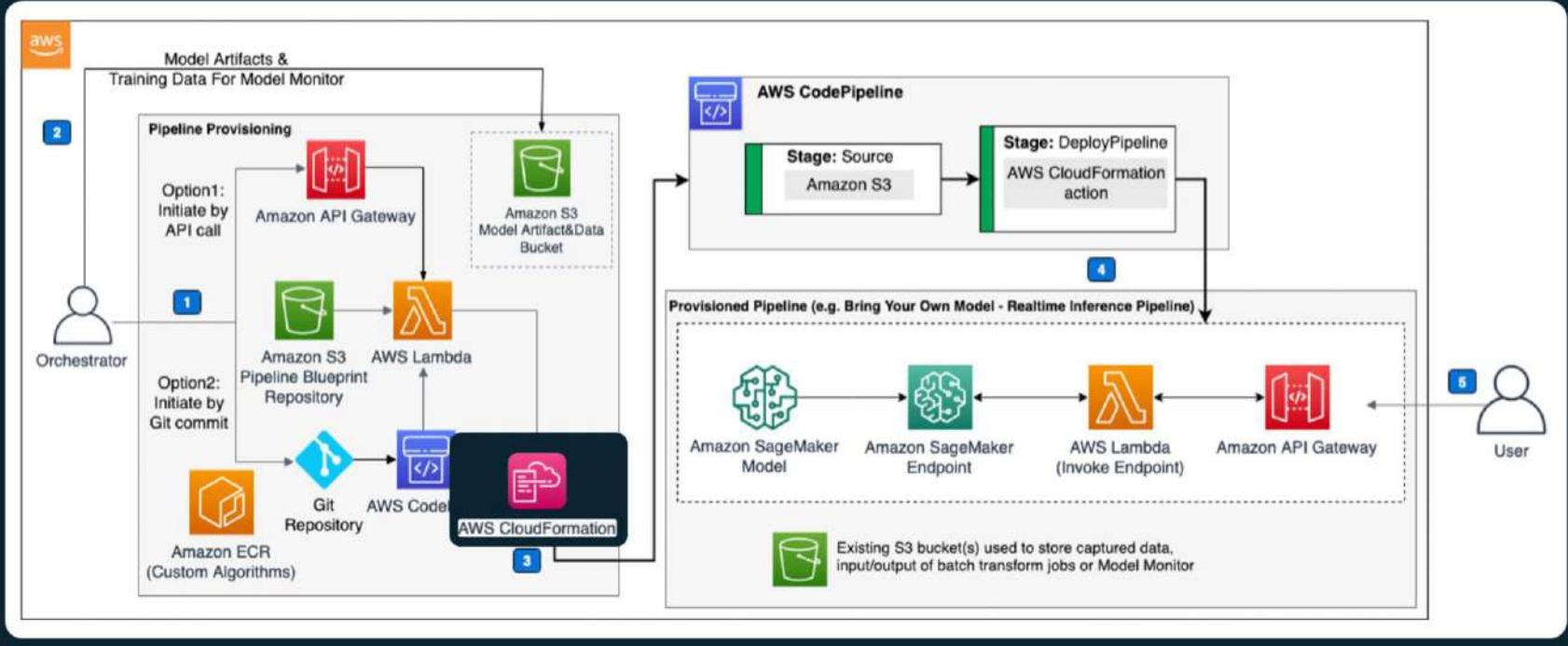


- Focus on experimentation 
- Smooth production integration 

© Copyright KodeKloud

With automated pipelines, data scientists can focus on experimentation and improving models while MLOps ensures smooth integration into production.

# Infrastructure as Code in MLOps



© Copyright KodeKloud

Source: <https://docs.aws.amazon.com/images/whitepapers/latest/ml-best-practices-public-sector-organizations/images/aws-mlops-framework.png>

Infrastructure as Code (IaC) allows the infrastructure required for machine learning pipelines, such as compute resources, databases, and networking components, to be defined and deployed via code. This makes ML environments easily reproducible and scalable. For instance, using tools like AWS CloudFormation or Terraform, data scientists can quickly spin up an ML environment to run experiments. This ensures the same infrastructure is used across different stages of model development, testing, and deployment, reducing errors and improving reliability.

Link: <https://docs.aws.amazon.com/images/whitepapers/latest/ml-best-practices-public-sector-organizations/images/aws-mlops-framework.png>

# Version Control in MLOps

Tracks code, data, models



Enables reverting and auditing



Ensures reproducibility



The screenshot shows the AWS CodeCommit interface. On the left, there's a sidebar with 'Developer Tools' and 'CodeCommit' selected. Under 'Source', 'Commits' is also selected. The main area is titled 'MyDemoRepo' and shows a 'Commit visualizer' tab selected. Below it, there are tabs for 'Commits' (selected), 'Commit visualizer' (highlighted in orange), and 'Compare commits'. The 'Commit visualizer' section displays a timeline of commits:

Commit ID	Message	Time Ago
d615e7ae	Merge branch 'AnotherBranch' into testbranch	2 minutes ago
b65b9863	Added another file.	2 minutes ago
73a6e39c	remote-tracking branch refs/remotes/origin/jane-branch into jane-branch	
6bbb6d3c	Another test of the editing feature.	20 minutes ago
edacdffe	Testing this out to see how well it works.	23 minutes ago
70bb94d7	Revised test results with correct information.	36 minutes ago
b78e6d1c	Merge branch 'Results' into testbranch	50 minutes ago
84b7d158	Edited ahs_count.py	50 minutes ago

In MLOps, version control is not only important for code but also for datasets, model configurations, and experiment results.

This enables teams to maintain lineage, track model changes, and revert to previous versions if needed. Tools like AWS Code Commit or GitLab are used to manage code, while model versions can be tracked in Amazon SageMaker Model Registry.

Version control ensures that every part of the ML lifecycle is reproducible, which is critical for auditing and improving the

models over time.

# Version Control in MLOps



AWS  
CodeCommit



GitLab

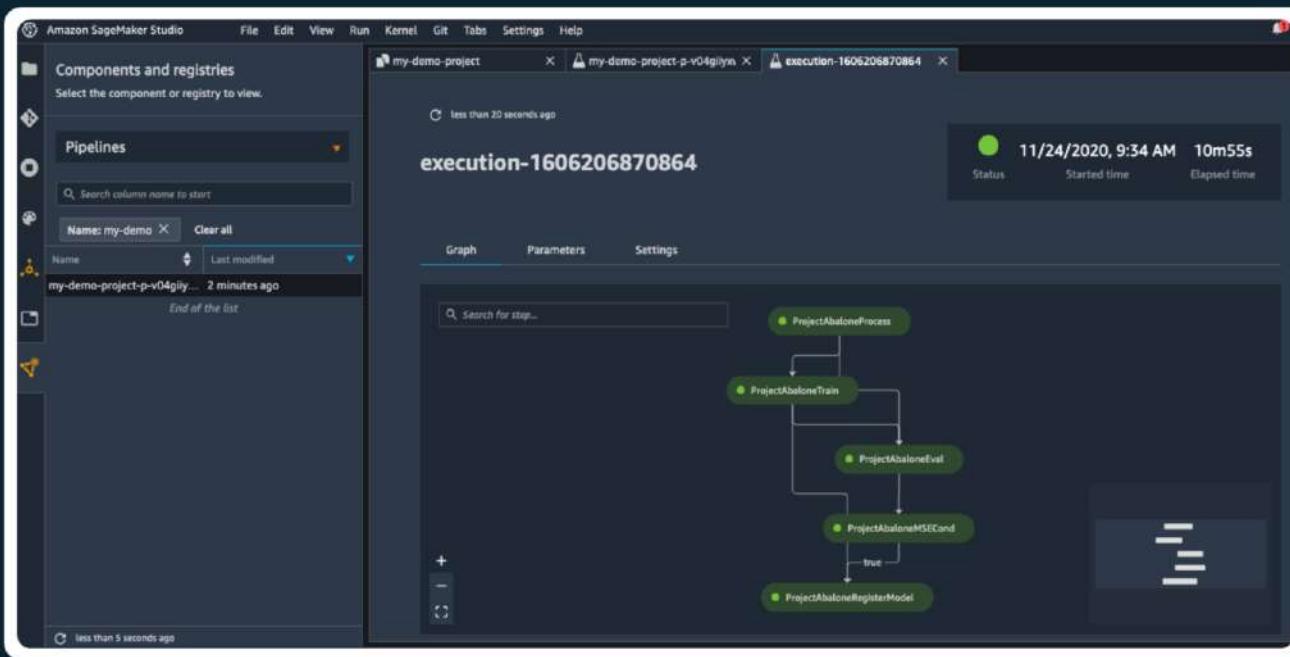


Amazon  
SageMaker

© Copyright KodeKloud

Tools like AWS Code Commit or GitLab are used to manage code, while model versions can be tracked in Amazon SageMaker Model Registry.

# Automating Model Training and Deployment



One of the key principles of MLOps is automation.

# Automating Model Training and Deployment

01



Reduces human intervention and errors

02



Automates data validation, model training, and tuning

03



Ensures seamless deployment of updated models

© Copyright KodeKloud

Automating tasks such as data validation, model training, and hyperparameter tuning reduces human intervention and errors.

Once models are trained, the deployment process is also automated, allowing for continuous integration of updated models into production environments.

# Automating Model Training and Deployment



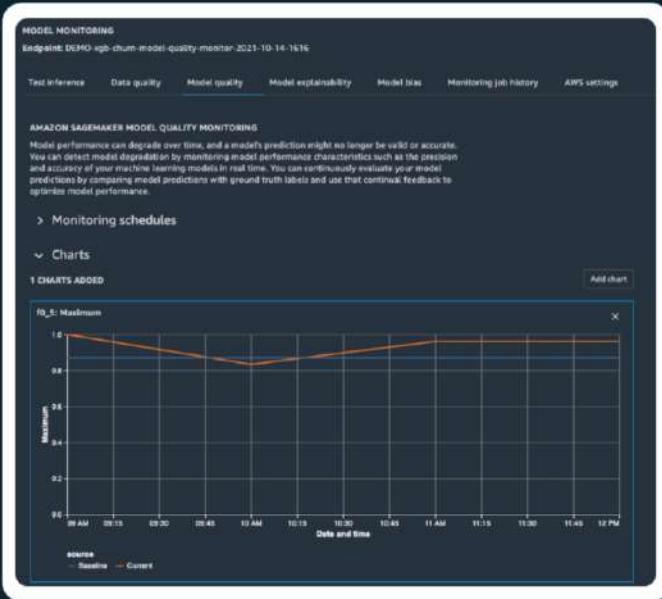
Amazon SageMaker

Automates **retraining** and **redeployment** as new data becomes available

© Copyright KodeKloud

For instance, Amazon SageMaker Pipelines can automate these steps, ensuring that models are retrained and redeployed seamlessly as new data becomes available.

# Monitoring and Retraining Models



Continuous monitoring ensures models perform as expected

Monitoring tools track accuracy, latency, and errors

Automatic retraining is triggered by performance degradation

Once deployed, models need continuous monitoring to ensure they are performing as expected. Changes in data distribution or external factors can lead to model drift, which impacts performance.

Monitoring tools track key metrics such as accuracy, latency, and errors. If performance degrades, MLOps pipelines can automatically trigger retraining of the models using new data.

# Monitoring and Retraining Models



Amazon CloudWatch



Amazon SageMaker

© Copyright KodeKloud

AWS offers services like Amazon CloudWatch and SageMaker Model Monitor to keep an eye on deployed models and automatically respond to issues.

# Compliance and Auditability in MLOps

```
{  
    "eventVersion": "1.05",  
    "userIdentity": {  
        "type": "IAMUser",  
        "principalId": "AIDAJQEXAMPLEUMLWJ6L",  
        "arn": "arn:aws:iam::123456789012:user/intern",  
        "accountId": "123456789012",  
        "accessKeyId": "ASXAIQEXAMPLEQULKNKV",  
        "userName": "intern"  
    },  
    "eventTime": "2018-01-02T15:23:46Z",  
    "eventSource": "sagemaker.amazonaws.com",  
    "eventName": "CreateModel",  
    "awsRegion": "us-west-2",  
    "sourceIPAddress": "127.0.0.1",  
    "userAgent": "USER_AGENT",  
    "requestParameters": {  
        "modelName": "ExampleModel",  
        "primaryContainer": {  
            "image": "174872318107.dkr.ecr.us-west-2.amazonaws.com/kmeans:latest"  
        },  
        "executionRoleArn": "arn:aws:iam::123456789012:role/EXAMPLEARN"  
    },  
    "responseElements": {  
        "modelArn": "arn:aws:sagemaker:us-west-2:123456789012:model/birminghamappy2018-01-02t15-23-32-275z-ivrdog"  
    },  
    "requestID": "417bd4b8-EXAMPLE",  
    "eventId": "6fb278e1-EXAMPLE",  
    "eventType": "AwsApiCall",  
    "recipientAccountId": "444455556666"  
}
```

# Compliance and Auditability in MLOps



Ensures documented and versioned ML lifecycle steps



Tracks model training, data usage, and deployment



Supports regulated industries like healthcare and finance



Demonstrates compliance with regulations

MLOps enhances compliance by ensuring that all steps in the machine learning lifecycle are documented and versioned. This includes tracking how models are trained, how data was used, and how the models were deployed. For industries like healthcare or finance, where audits are common, MLOps ensures that organizations can demonstrate compliance with regulations.

# Compliance and Auditability in MLOps



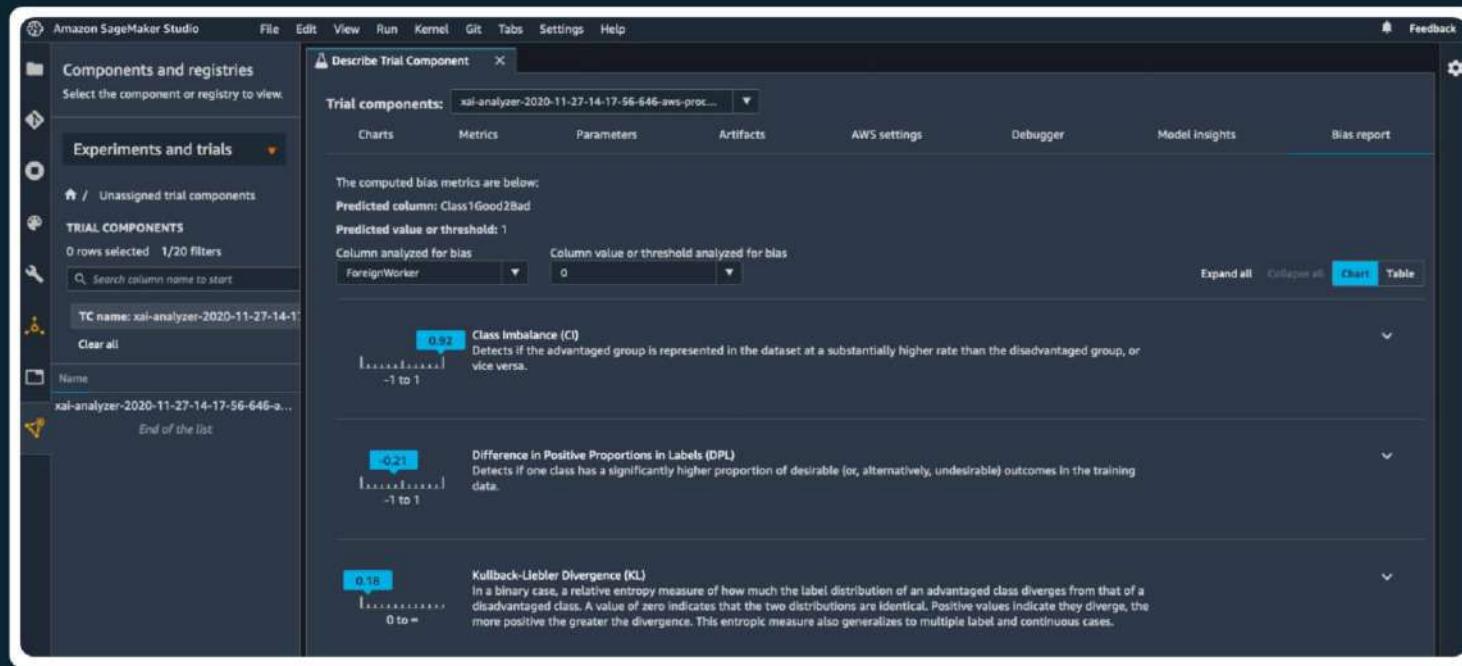
**Amazon SageMaker**

**Logs** and **audits** the model training and deployment process for review

© Copyright KodeKloud

Amazon SageMaker offers built-in tools for logging and auditing the entire model training and deployment process, allowing organizations to review exactly how a model was created and deployed.

# Improving Model Quality With MLOps



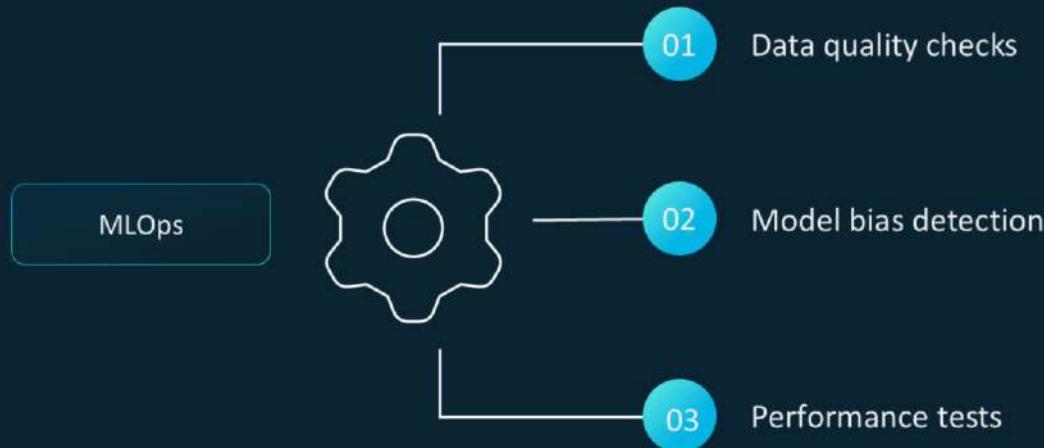
© Copyright KodeKloud

Source: <https://noise.getoto.net/author/julien-simon/>

MLOps provides mechanisms for ensuring models meet quality standards, including tracking performance metrics and enforcing policies to reduce bias or improve fairness.

Regular evaluations are essential to maintaining the quality of models in production.

# Improving Model Quality With MLOps



. MLOps pipelines ensure that data quality checks, model bias detection, and performance tests are automated.

# Improving Model Quality With MLOps



Amazon SageMaker Clarify

**Monitors fairness and bias**, ensuring models are accurate and equitable

© Copyright KodeKloud

Tools like Amazon SageMaker Clarify can help monitor for fairness and bias, ensuring models are both accurate and equitable.

# MLOps Tools – Amazon SageMaker Pipelines

01



End-to-end  
automation

02



Flexible pipeline  
definition

03



Workflow visualization

04



Seamless  
integration

© Copyright KodeKloud

Amazon SageMaker Pipelines is a managed service that allows users to create end-to-end machine learning pipelines. With SageMaker Pipelines, data scientists can automate the process of building, training, and deploying models. The entire pipeline can be defined in Python or JSON, and users can visualize workflows in SageMaker Studio. SageMaker Pipelines also integrates with other AWS services like CodeCommit for version control and CloudWatch for monitoring, providing a complete MLOps solution.

# Evaluating ML Models – Confusion Matrix

		Actual Value	
		Positive	Negative
Predicted Value	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Common tool for evaluating classification models

Summarizes model predictions vs actual outcomes

Identifies errors: False positives and false negatives

Derives metrics such as accuracy, precision, and recall

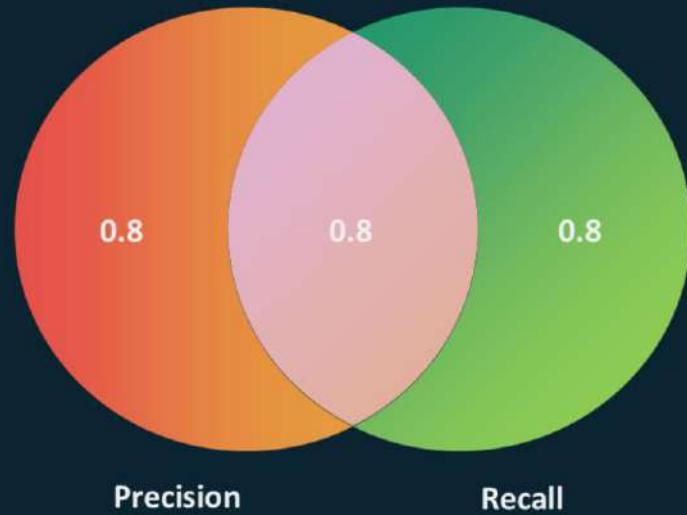
© Copyright KodeKloud

A confusion matrix is a common tool for evaluating classification models, summarizing how well the model's predictions match actual outcomes.

It helps in identifying errors, such as false positives (incorrectly predicting a positive outcome) and false negatives (failing to predict a positive outcome when it should have).

Metrics like accuracy, precision, and recall are derived from the confusion matrix, giving a comprehensive picture of model performance.

# Model Performance Metrics – Precision, Recall, and F1 Score



01

## Precision

Measures correct positive predictions; minimizes false positives  
(e.g., spam detection)

Precision measures how many of the positive predictions were correct, making it useful in scenarios where minimizing false positives is critical, such as in spam detection.

# Model Performance Metrics – Precision, Recall, and F1 Score



02

## Recall

Captures actual positives; vital for medical diagnoses

Recall, also known as sensitivity, measures how well the model captures actual positives, useful in contexts like medical diagnoses where missing a positive case could be costly.

# Model Performance Metrics – Precision, Recall, and F1 Score



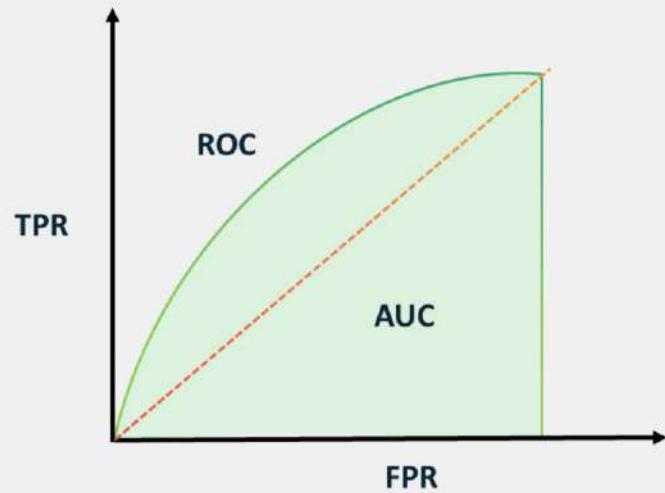
03

## F1 Score

Balances precision and recall; useful for false positive/ negative trade-offs

The F1 score balances precision and recall, offering a single metric that accounts for both. This is especially valuable when there is a need to balance between false positives and false negatives.

# Area Under Curve (AUC) for Binary Classification



## Area Under the Curve (AUC)

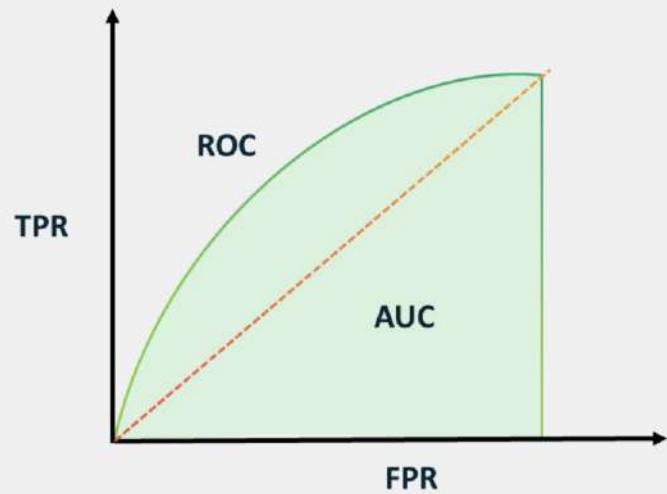
- Evaluates binary classification model performance
- Ranges from **0.5** (random guessing) to **1** (perfect prediction)

© Copyright KodeKloud

The Area Under the Curve (AUC) is used to evaluate the performance of binary classification models. AUC values range from 0.5 (random guessing) to 1 (perfect prediction). A higher AUC score indicates better model performance.

It is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various thresholds.

# Area Under Curve (AUC) for Binary Classification



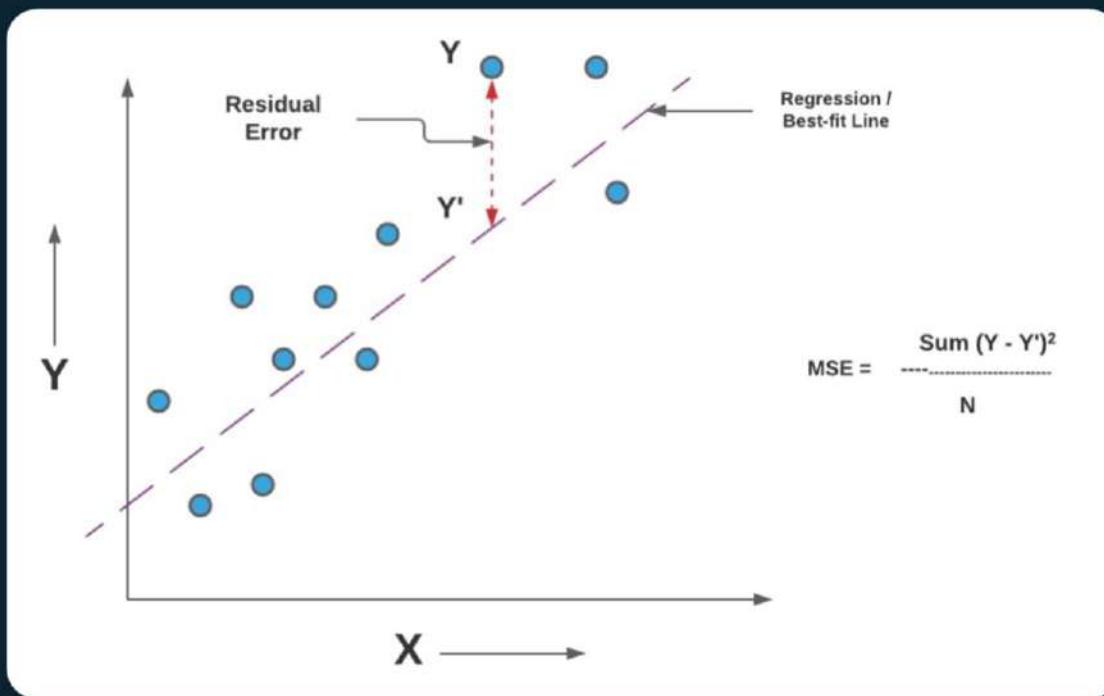
## Receiver Operating Characteristic (ROC)

Plots true positive rate vs false positive rate at various thresholds

© Copyright KodeKloud

It is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various thresholds.

# Mean Squared Error (MSE) in Regression Models



# Mean Squared Error (MSE) in Regression Models

01

Evaluates regression models

02

Calculates the average of the squares of errors

03

Indicates that a smaller MSE means better predictions

04

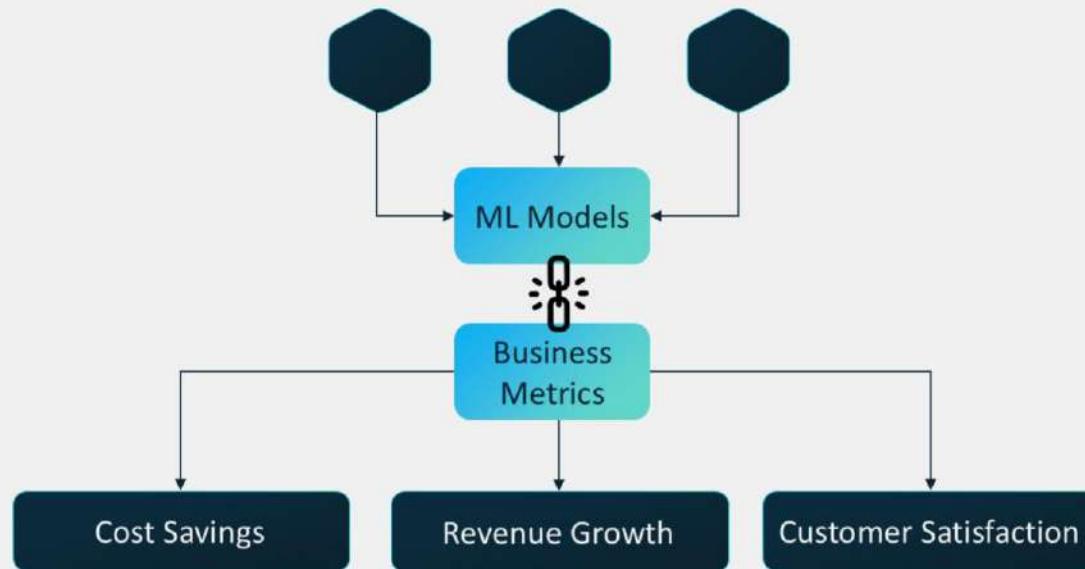
Emphasizes large errors and is sensitive to outliers

Mean Squared Error (MSE) is commonly used to evaluate regression models. It calculates the average of the squares of the errors, providing a measure of how well the model's predictions match actual outcomes.

The smaller the MSE, the better the model is at making accurate predictions. However, MSE emphasizes large errors, making it sensitive to outliers.

The Root Mean Squared Error (RMSE) is often used because it presents errors in the same units as the dependent variable, making it easier to interpret.

# Business Metrics and ROI for ML Projects



**Setting business goals early ensures model alignment with organizational priorities.**

© Copyright KodeKloud

It's crucial to connect machine learning models to business metrics, such as cost savings, revenue growth, or customer satisfaction improvements.

Defining business goals at the start of the ML lifecycle ensures that models align with organizational priorities.

# Business Metrics and ROI for ML Projects



To assess the return on investment (ROI) for ML projects

© Copyright KodeKloud

After deployment, actual performance is tracked and compared to the initial goals, allowing businesses to calculate the return on investment (ROI) of their ML projects.

# AWS Tools for MLOps



Amazon  
SageMaker



AWS  
CodeCommit



AWS Step  
Functions



Amazon  
CloudWatch



Amazon  
SageMaker  
Model Monitor

© Copyright KodeKloud

AWS offers a range of tools to support MLOps, including:

Amazon SageMaker for building, training, and deploying models.

AWS CodeCommit for version control of code and models.

AWS Step Functions for orchestrating serverless workflows.

Amazon CloudWatch for monitoring deployed models.

Amazon SageMaker Model Monitor for tracking model performance over time.

These tools create a seamless integration between development, operations, and monitoring, making MLOps a powerful solution for scaling ML workflows.



KodeKloud

# AI and ML Services on AWS

---

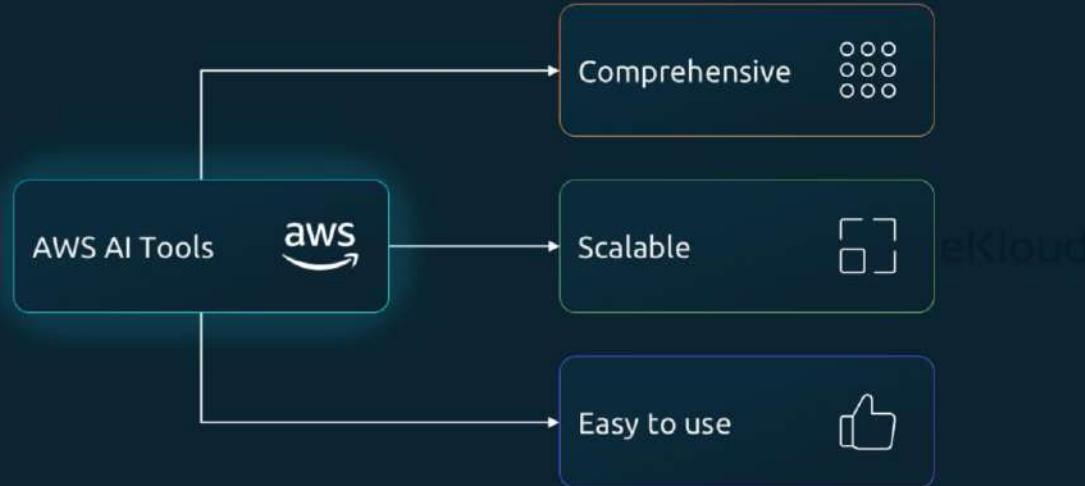
Overview



# AWS for AI



# AWS Is a Leader in AI Services



© Copyright KodeKloud

AWS is a cloud leader, providing comprehensive, scalable, and easy-to-use AI tools that allow businesses to implement AI without needing deep expertise in the field.

# AWS Is a Leader in AI Services



**AWS SageMaker** – Simplify Machine Learning development



**AWS Polly** – Pre-trained model to enable quick AI integration



**AWS Lex** – Pre-trained model to enable quick AI integration



Services like Amazon SageMaker simplify machine learning development, and pre-trained models such as Amazon Polly and Lex enable developers to integrate AI into their applications quickly.

# AWS Supports Scalable AI Solutions

01



Global Scalability

02



Pay-as-You-Go

03



High Availability

04



Security Features

AWS allows businesses to scale their AI solutions globally, with the flexibility to only pay for what they use. Furthermore, AWS's high availability and security features make it a trusted choice for enterprise-level AI solutions.

# AWS AI/ML – Service Categories

01

Pre-trained AI services

- Ready-to-use APIs
- No ML expertise needed

02

ML platforms

- Tools for custom ML model development

03

Infrastructure for custom AI

- Training infrastructure, including GPU instances

**Pre-trained AI Services:** These services provide ready-to-use APIs that developers can integrate into their applications. No machine learning expertise is needed.

**ML Platforms:** AWS offers comprehensive tools for developing custom ML models, from data preparation to model training and deployment.

**Infrastructure for Custom AI:** AWS also offers the infrastructure to train deep learning models, like GPU instances and distributed training environments.

# AWS Bedrock



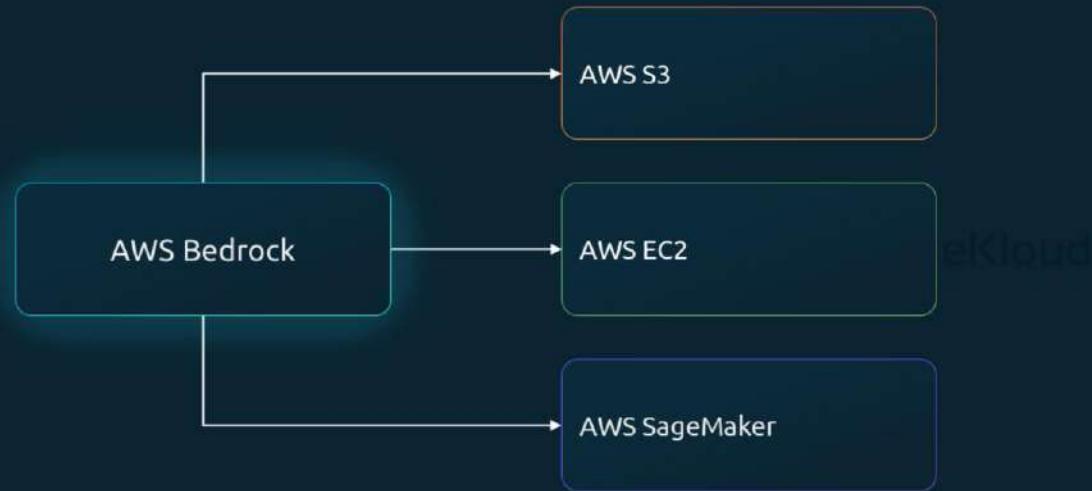
# AWS Bedrock



## What Is AWS Bedrock?

**Amazon Bedrock is a fully managed service for foundation models that simplifies Generative AI application development.**

# AWS Bedrock – Capabilities



© Copyright KodeKloud

Developers can choose from a variety of pre-trained models.  
You can customize these models with your own data securely.  
Integration with other AWS services like S3, EC2, and SageMaker is straightforward.

# AWS Bedrock – Capabilities

Access to diverse foundation models

Easy customization with your data

Seamless integration with AWS services

Developers can choose from a variety of pre-trained models.  
You can customize these models with your own data securely.  
Integration with other AWS services like S3, EC2, and SageMaker is straightforward.

# AWS Bedrock – Guardrails

Ensuring safe and responsible AI

Built-in content moderation

Compliance and data privacy features

Guardrails help ensure that AI applications produce safe and appropriate content. AWS Bedrock includes tools for content moderation to filter out harmful or sensitive information. It assists in maintaining compliance with data privacy regulations like GDPR.

# AWS Bedrock Agents

Automate complex tasks

Orchestrate workflows based on AI outputs

Enhance application capabilities

Bedrock Agents are tools that automate tasks and workflows.  
They can perform actions based on AI outputs, like retrieving data or triggering processes.  
This enhances the functionality of AI applications, making them more interactive and dynamic.

# **Amazon SageMaker –**

## Build, Train, and Deploy ML Models



Amazon SageMaker

**A managed service for quickly building, training, and  
deploying ML models, simplifying development**

# SageMaker – Overview

**01**



Data preparation

**02**



Model training

**03**



Tuning

**04**



Deployment

Overview of Amazon SageMaker: Amazon SageMaker is a fully managed service that enables developers to build, train, and deploy ML models quickly and efficiently. It removes the heavy lifting required in traditional machine learning development.

# Key Features

**01**



Data labeling tools

**02**



Notebooks for model development

**03**



Automatic model tuning

**04**



Scalable deployment

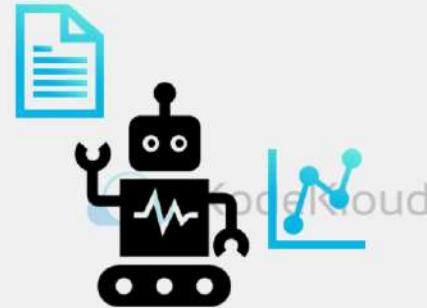
**05**



Easy-to-use for those with minimal AI expertise

Key SageMaker features: SageMaker includes modules for data labeling, notebooks for model development, automatic model tuning, and deployment at scale. With SageMaker, even those with minimal AI expertise can develop ML models.

# SageMaker



© Copyright KodeKloud

## Machine Learning

Machine learning (ML) involves algorithms that detect patterns in data to make predictions or decisions, while human learning is rooted in experience, observation, and reasoning.

# SageMaker Workflow



© Copyright KodeKloud

## Data Ingestion:

Amazon S3: In the context of SageMaker, raw data is usually stored in Amazon S3 buckets, which serve as the primary data storage solution.

## Data Preparation & Exploration:

SageMaker Notebook: Data scientists use Jupyter notebooks provided by SageMaker to explore, clean, and preprocess the

data. This step might involve visualizing the data, handling missing values, and feature engineering.

**SageMaker Data Wrangler:** This is a tool within SageMaker that simplifies the process of data preparation and feature engineering

#### Model Training:

**Training Data:** A subset of the preprocessed data is used to train the machine learning model.

**SageMaker Training Jobs:** SageMaker provisions the required infrastructure to train the model. You can specify the type and number of instances you need.

#### Model Evaluation & Tuning:

**Validation Data:** Another subset of the preprocessed data is used to validate the model's performance.

**Automatic Model Tuning:** SageMaker can automatically tune model hyperparameters to optimize performance.

#### Model Deployment:

**SageMaker Endpoints:** Once the model is trained and tuned, it can be deployed to a SageMaker endpoint for real-time predictions.

# SageMaker



© Copyright KodeKloud

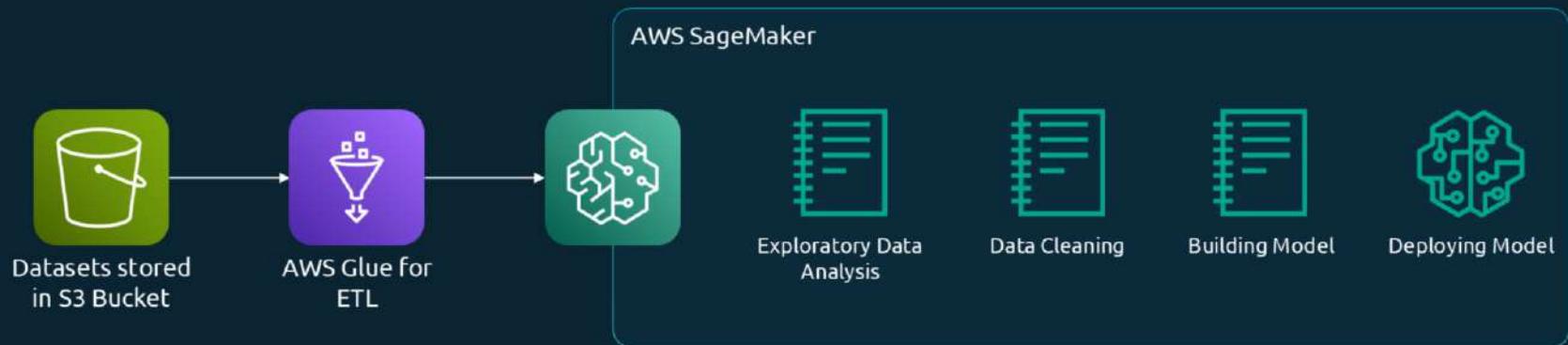
**SageMaker Notebook Instances:** These are fully managed ML compute instances running the Jupyter Notebook App. They allow data scientists to quickly spin up notebook environments to experiment and build models.

**SageMaker Training Jobs:** This is where the actual training of the machine learning models happens. You can specify the type and number of instances you need, and SageMaker manages the rest.

SageMaker Models: Machine learning models that can be hosted on Amazon SageMaker from job output

SageMaker Endpoints: After you've trained a model, you can deploy it using SageMaker Endpoints, which allows you to make real-time predictions.

# SageMaker



© Copyright KodeKloud

**AWS Glue for ETL:** Data from the S3 bucket is processed using AWS Glue, an ETL (Extract, Transform, Load) service. This prepares the data for analysis.

**AWS SageMaker:** Once the data is processed and transformed, it's fed into AWS SageMaker for various stages of machine learning:

Exploratory Data Analysis: This is the initial phase where the data is inspected to understand its structure, anomalies, patterns, and relationships.

**Data Cleaning:** Any anomalies, missing values, or discrepancies detected in the previous phase are addressed in this phase to ensure the quality of the dataset.

**Building Model:** With the cleaned data, a machine learning model is trained to make predictions or classifications.

**Deploy Model:** Once the model is built and trained, it is deployed to be used in real-world scenarios.

# Features

01



Built-in algorithms  
and BYOA

02



Integrated Jupyter  
Notebooks

03



Distributed training

04



Automatic model  
tuning

05



SageMaker studio

© Copyright KodeKloud

**Built-in Algorithms and BYOA:** SageMaker provides a set of high-performance, scalable machine learning algorithms optimized for speed. These cover a wide range of use cases, from regression to classification and clustering. It also allows you to bring your own custom algorithms or use third-party libraries.

**Integrated Jupyter Notebooks:** Provides built-in Jupyter notebooks for easy data exploration, cleaning, and preprocessing.

Distributed Training: SageMaker supports distributed training on multiple instances, allowing you to train large models on vast datasets faster. It uses Amazon EC2 Spot instances for training jobs, reducing costs.

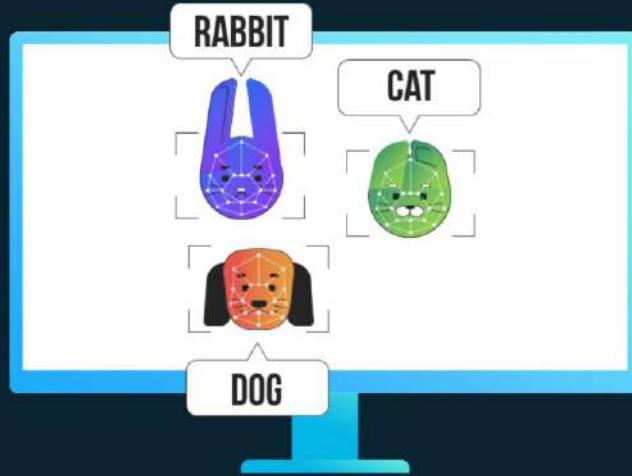
Automatic Model Tuning: Uses machine learning to automatically search and optimize for the best hyperparameters.

SageMaker Studio: An integrated development environment (IDE) for machine learning, providing a single, web-based visual interface for the complete ML workflow.

# **Amazon Rekognition –** Image and Video Analysis



# Introduction



© Copyright KodeKloud

What is Amazon Rekognition? Amazon Rekognition is a service that enables developers to analyze images and videos using pre-trained models for tasks such as object detection, facial analysis, and text recognition.

## What Is Amazon Rekognition?

**Amazon Rekognition analyzes images and videos for object detection, facial analysis, and text recognition.**

© Copyright KodeKloud

What is Amazon Rekognition? Amazon Rekognition is a service that enables developers to analyze images and videos using pre-trained models for tasks such as object detection, facial analysis, and text recognition.

# Use Cases

Facial recognition in security systems

Analyzing visual content for media companies

Recognizing logos or product features in retail

Use cases for Amazon Rekognition: It can be used in security systems for facial recognition, in media companies for analyzing large volumes of visual content, or in retail for recognizing logos or product features.

# Rekognition



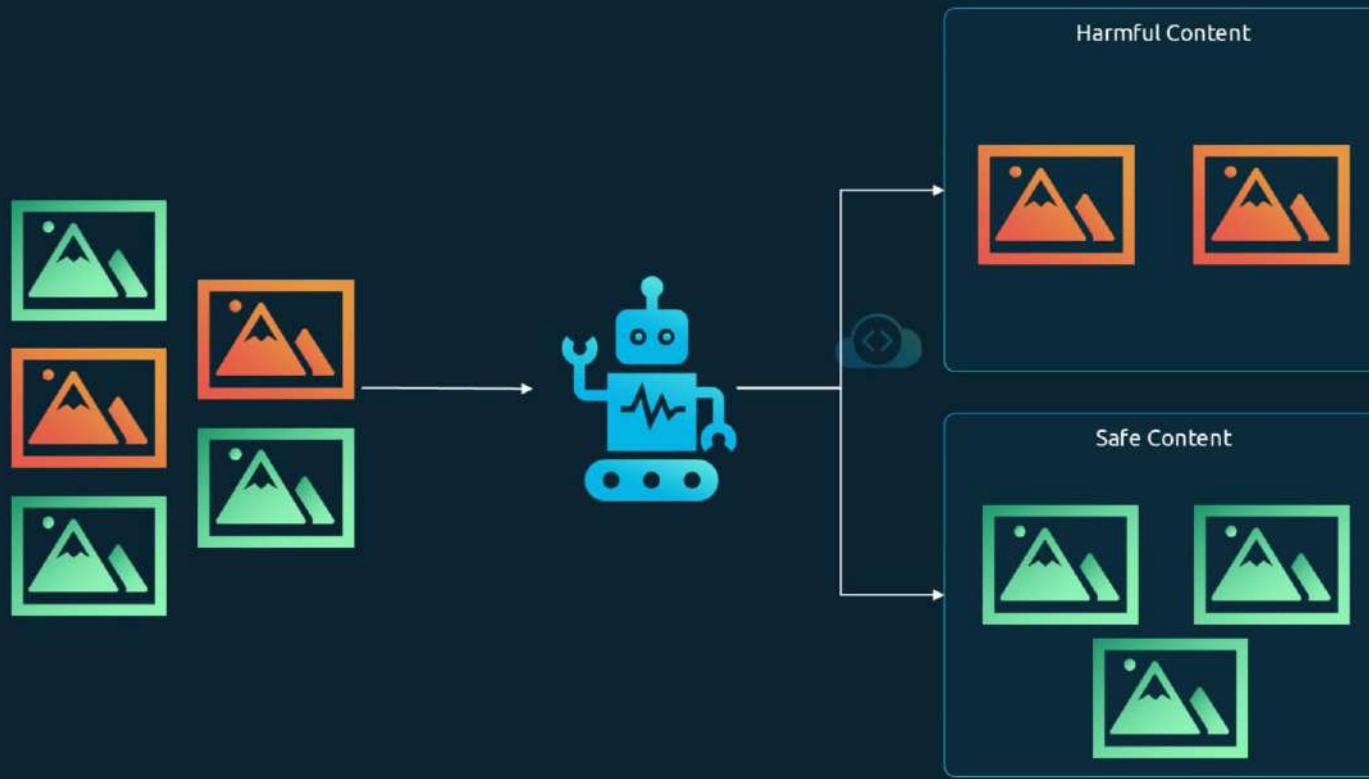
© Copyright KodeKloud

An automated tagging system where, instead of manually tagging each image, Amazon Rekognition provides relevant and detailed tags based on the image's content. This can be particularly useful for large image databases, helping in image categorization, search, and management.

Amazon Rekognition is a service offered by Amazon Web Services (AWS) that uses deep learning technology to analyze and interpret images and videos. It's designed to recognize objects, people, text, scenes, and activities in images and

videos, as well as detect any inappropriate content.

# Rekognition



© Copyright KodeKloud

Content moderation is the process of monitoring, reviewing, and managing user-generated content (UGC) to ensure it adheres to established guidelines and standards. This process is crucial for maintaining the quality and safety of online platforms, especially those that rely heavily on user contributions.

# Rekognition



© Copyright KodeKloud

**Amazon S3:** Amazon Rekognition can directly analyze images and videos stored in Amazon Simple Storage Service (S3) buckets. This integration simplifies the process of accessing and processing large amounts of data.

**AWS Lambda:** You can use AWS Lambda to trigger functions in response to Amazon Rekognition events. For example, when a face is detected in a video, a Lambda function can be triggered to take specific actions.

**Amazon DynamoDB:** Results from Amazon Rekognition can be stored in DynamoDB tables. For instance, metadata from recognized faces can be stored for future reference or analysis.

**Amazon Augmented AI (A2I):** For use cases that require human review of the predictions made by Amazon Rekognition, you can integrate with Amazon A2I to easily implement human review workflows.

# Features

01



Object and scene  
detection

02



Facial analysis and  
recognition

03



Text in image

04



Activity detection

05



Unsafe content  
detection

© Copyright KodeKloud

Object and Scene Detection: Amazon Rekognition can identify thousands of objects such as vehicles, pets, or furniture, and scenes like sunsets, beaches, or weddings.

Facial Analysis and Recognition: The service can detect faces within images and videos, identify specific facial features, provide sentiment analysis, and (with proper legal and ethical considerations) match faces against databases to recognize individuals.

**Text in Image:** Amazon Rekognition can detect and recognize text within images, such as street names, captions, product names, and license plates.

**Activity Detection:** In video files, Amazon Rekognition can detect activities, such as "running," "jumping," "eating," etc.

**Unsafe Content Detection:** Amazon Rekognition can identify potentially unsafe or inappropriate content across both image and video files, making it easier to filter content for your application.

# Features

06



Celebrity  
recognition

07



Custom labels

08



Integration with  
other AWS services

09



Emotion detection

10



Real-time analysis

Celebrity Recognition: The service can recognize celebrities in images and videos from a wide array of fields such as politics, sports, business, and entertainment.

Person Tracking: Amazon Rekognition makes it possible to automatically follow the path of a person in a stored video. It can also help in counting people.

**Face Analysis:** The service provides detailed facial analysis, including attributes such as estimated age range, gender, whether a person is smiling, if they're wearing glasses, and the position and quality of the face.

**Custom Labels:** With Amazon Rekognition Custom Labels, you can identify the objects and scenes in images that are specific to your business needs by using a custom machine learning model.

**Integration with other AWS services:** Amazon Rekognition integrates with other AWS services like Amazon S3, AWS Lambda, and Amazon Augmented AI (A2I) for human review of automated predictions.

**Real-Time Analysis:** Amazon Rekognition can analyze live video streams in real time to detect and recognize faces, identify objects, and flag inappropriate content.

**Emotion Detection:** It can analyze faces to detect emotions such as happiness, sadness, anger, surprise, disgust, calmness, and confusion

# **Amazon Lex –** Conversational AI



# Introduction



© Copyright KodeKloud

What is Amazon Lex? Amazon Lex is a service that allows you to build conversational interfaces, like chatbots and virtual assistants, using voice and text. It powers Amazon Alexa and allows developers to create AI-driven customer support systems.

## What Is Amazon Lex?

**Amazon Lex enables building chatbots and virtual assistants using voice and text. It powers Amazon Alexa and supports AI-driven customer support.**

What is Amazon Lex? Amazon Lex is a service that allows you to build conversational interfaces, like chatbots and virtual assistants, using voice and text. It powers Amazon Alexa and allows developers to create AI-driven customer support systems.

# Use Cases

- Customer service bots
- Automated call centers
- Virtual assistants for booking systems, FAQs, or troubleshooting guides

Use cases for Amazon Lex: Common applications include customer service bots, automated call centers, and virtual assistants that can handle booking systems, FAQs, or troubleshooting guides.

# Lex



© Copyright KodeKloud

Hotel booking system, where the chatbot assists users in making reservations by gathering necessary details interactively (user-to-chatbot interaction)

# Lex



© Copyright KodeKloud

**Amazon Cognito:** This service is primarily used for user authentication and identity management.

**Amazon Lex:** Once authenticated, the user interacts with an application powered by Amazon Lex.

**AWS Lambda:** After Lex understands the intent of the user's input, it can trigger an AWS Lambda function. In this case, it can process the user's request, potentially fetching, modifying, or storing information.

**Amazon DocumentDB:** Lambda interfaces with Amazon DocumentDB, which is a managed MongoDB-compatible database service. Lambda can retrieve data from DocumentDB, update records, or store new information based on the user's

interaction with Lex.

Imagine a user logs into a support chatbot using their credentials (managed by Cognito). They ask the chatbot (Lex) about the hotel booking. Lex, after understanding the intent, triggers a Lambda function which queries the user's hotel booking status from DocumentDB. Lambda processes the data, and Lex then responds to the user with the appropriate information.

# Features

01



Natural Language  
Understanding (NLU)  
and Automatic Speech  
Recognition (ASR)

02



Easy to build

03



Fully managed

04



Built-in integrations

05



Multi-channel support

© Copyright KodeKloud

Natural Language Understanding (NLU) and Automatic Speech Recognition (ASR):

Amazon Lex interprets user input using automatic speech recognition (ASR) to convert speech to text, and natural language understanding (NLU) to recognize the intent of the text. This enables you to build applications with highly engaging user experiences and lifelike conversational interactions.

Easy to Build:

You can create your own conversational bot using the Amazon Lex console, and define the conversation flow with an intuitive interface.

**Fully Managed:**

Amazon Lex is a fully managed service so you don't need to worry about managing infrastructure. Scaling is handled automatically, so your bot can handle peak periods.

**Built-in Integrations:** Amazon Lex integrates with AWS Lambda for executing logic, Amazon Cognito for user authentication, and Amazon Polly for text-to-speech capabilities.

**Multi-Channel Support:**

You can publish your Amazon Lex bot on mobile devices, web applications, chat services, and IoT devices, among others. Amazon Lex supports multiple messaging platforms and allows seamless migration between different channels.

# Lex



© Copyright KodeKloud

**Voice Assistant:** Similar to Siri or Alexa, you can create a custom voice assistant for specific tasks. The user speaks a command, Amazon Lex processes and understands the intent, and then AWS Polly vocalizes the response.

**Customer Support Chatbot:** On a website or mobile application, a chatbot can assist users. They type in or speak their questions, Amazon Lex provides the answers, and if the interaction is voice-based, AWS Polly can read the answers out loud.

# Amazon Polly

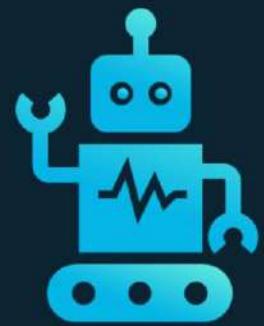
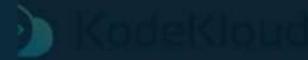


# Polly



Hey, tell me  
about the  
weather

and sunny.  
is mostly  
ith high of 35  
C and low of 27 C



© Copyright KodeKloud

Voice-based weather bot



© Copyright KodeKloud

### Voice-based weather bot:

A user speaks a request to know about the weather, the system fetches the relevant weather information from a Weather API, and then Amazon Polly converts the text-based weather data into a spoken response, which is relayed back to the user through the voice assistant.

# Features

01



Lifelike speech

02



Real-time streaming  
or file generation

03



SSML support

04



Lexicon support

05



Integration with  
other AWS services

**Lifelike Speech:** Amazon Polly uses advanced deep learning technologies to synthesize speech that sounds like a human voice. It includes dozens of lifelike voices across a variety of languages, so you can select the ideal voice and build speech-enabled applications that work in many different countries.

**Real-time Streaming or Stored Audio:** Amazon Polly lets you stream synthesized speech to your application in real-time, allowing you to play it directly with no delay. Alternatively, you can also store synthesized speech in standard audio file

formats and distribute it without further processing.

**SSML Support:** Amazon Polly supports Speech Synthesis Markup Language (SSML), which allows you to control aspects of speech such as pronunciation, volume, pitch, speed rate, etc. This makes the speech output more natural and expressive.

**Lexicon Support:** You can customize the pronunciation of words by defining a lexicon. This feature allows you to control the pronunciation of words or phrases specific to your business or use case.

**Integration with Other AWS Services:** Amazon Polly integrates with other AWS services like Amazon Lex for building conversational interfaces, AWS Lambda for serverless computing, or Amazon S3 for storing the generated speech files.

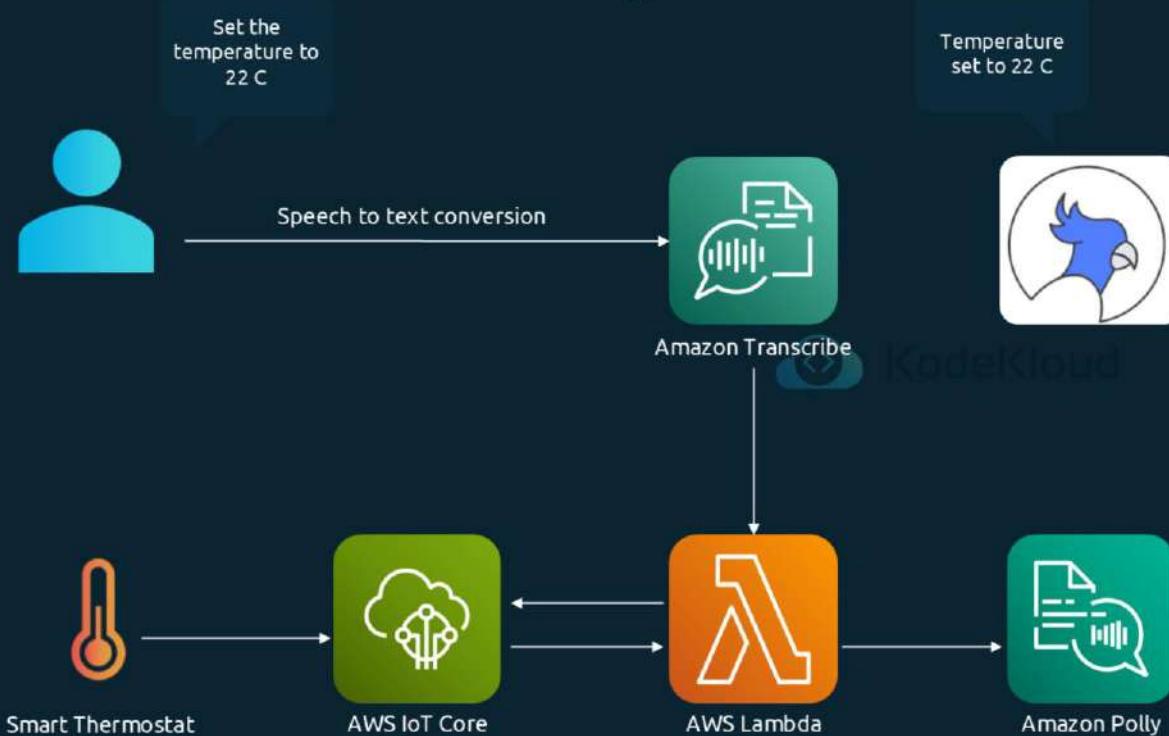
# Polly



© Copyright KodeKloud

A transcription file is uploaded to an Input S3 Bucket.  
The upload action triggers an AWS Lambda function.  
The Lambda function processes the uploaded text and sends it to Amazon Polly.  
Amazon Polly converts the text into speech, producing an audio file.  
The generated audio file is then saved to an Output S3 Bucket.

# Polly



© Copyright KodeKloud

In this example, a smart thermostat is equipped with sensors to detect room temperature and is configured on AWS IoT Core. Communication is established using the MQTT protocol.

AWS Transcribe is the service used to convert speech into text.

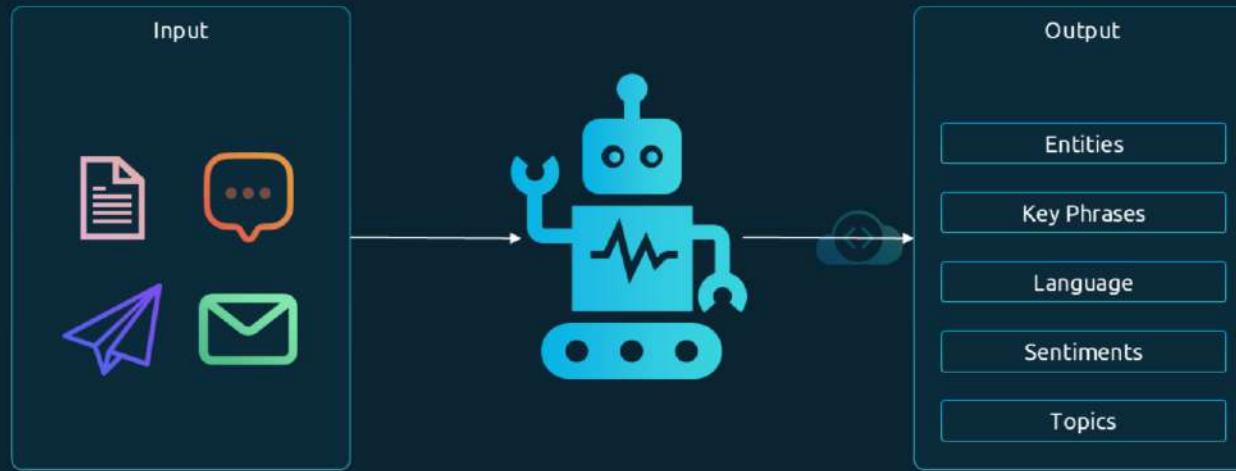
Upon receiving a voice command, an AWS Lambda function is triggered. It sends a request to AWS IoT Core to set the

desired temperature. Once this action is completed, the Lambda function sends a request to Amazon Polly with the text "Temperature set to 22 degrees" to convert the confirmation message into speech.

# Amazon Comprehend



# Comprehend



© Copyright KodeKloud

Input: Social media posts, emails, documents, transcription etc

Output:

Entities: This refers to named entities that AWS Comprehend can identify within the text. Examples include people's names, dates, locations, organizations, brands, monetary values, and more. For instance, in the sentence "Apple Inc. was founded in April 1976," "Apple Inc." and "April 1976" would be identified as entities.

**Key Phrases:** Comprehend can detect and extract key phrases or important terms within a given text. For example, in the sentence "Cloud computing offers scalable resources," "Cloud computing" and "scalable resources" would be highlighted as key phrases.

**Language:** AWS Comprehend can automatically detect the dominant language in which a text is written. For texts containing multiple languages, Comprehend will identify the primary language being used.

**Sentiments:** The service can classify the sentiment of a piece of text. Typical sentiment outputs include "Positive", "Negative", "Neutral", and "Mixed". For instance, the text "I love using AWS services" would likely be categorized as "Positive".

**Topics:** In the context of topic modeling, AWS Comprehend can analyze a collection of documents and identify common themes or topics within them. This is particularly useful for understanding large volumes of unstructured text data. For example, if analyzing a set of news articles, topics might include "technology", "politics", "sports", etc.

# Features

01



Language detection

02



Sentiment analysis

03



Entity recognition

04



Key Phrases  
extraction

05



Topic modelling

© Copyright KodeKloud

Language Detection: Amazon Comprehend can detect the dominant language from a text and supports multiple languages.

Sentiment Analysis: This feature determines whether the text is positive, negative, neutral, or mixed. This is particularly useful for understanding customer feedback.

**Entity Recognition:** Amazon Comprehend identifies various types of entities such as names of people, cities, organizations, dates, quantities, and more.

**Key Phrases Extraction:** It extracts key phrases from the text, helping you understand the main points.

**Topic Modeling:** Amazon Comprehend can analyze a collection of documents and divide them into topics, which can help in content organization, document filtering, and more.

# Features

06



Syntax analysis

07



PII detection

08



Multilingual support

09



Integration with  
other AWS services

10



Real-time processing

**Syntax Analysis:** Using tokenization and parts of speech, it can understand the roles and relationships of words in a sentence.

**PII Detection:** Amazon Comprehend can also identify personally identifiable information (PII) in a text, helping you protect customer data.

**Multi-lingual Support:** Amazon Comprehend supports multiple languages, allowing businesses to analyze and understand documents from diverse demographics.

**Integration with other AWS Services:** It integrates with other AWS services like Amazon S3, AWS Glue, and AWS Lambda, allowing you to analyze data in various formats and sources easily.

**Real-time Processing:** Amazon Comprehend can process documents in real-time, providing immediate insights for applications like social media monitoring, data analytics, and more.

# Comprehend



© Copyright KodeKloud

Sentiment analysis for Product reviews

# Comprehend



© Copyright KodeKloud

Sentiment analysis for Product reviews:

Amazon S3: These reviews are then uploaded or stored in Amazon S3. AWS Lambda: Once the reviews are in S3, an AWS Lambda function is triggered. In this context, the Lambda function could be responsible for preprocessing the reviews or invoking other services, such as Amazon Comprehend, for analysis.

Amazon Comprehend: Lambda then passes the reviews to Amazon Comprehend, a natural language processing (NLP)

service. Amazon Comprehend will analyze the reviews for sentiment (e.g., positive, negative, neutral, or mixed). This service uses machine learning to detect the sentiment and can return detailed sentiment scores for each review.

Athena: Alongside sentiment analysis, the reviews stored in Amazon S3 can be queried using Amazon Athena. For instance, you might use Athena to aggregate the reviews, filter them based on certain criteria, or prepare them for visualization.

Amazon QuickSight: Finally, the sentiment analysis results and any data processed by Athena can be visualized using Amazon QuickSight. For instance, you might visualize the distribution of positive versus negative reviews over time or compare sentiment across different product categories.

# Amazon Fraud Detector



# Fraud Detector



© Copyright KodeKloud

Amazon Fraud Detector is a fully managed service that enables users to identify potentially fraudulent activities more effectively.

It allows users to build, deploy, and manage fraud detection models without prior machine learning experience.

# Fraud Detector



© Copyright KodeKloud

Here Fraud detector helps in reducing online payment fraud by identifying potentially suspicious transactions before processing payments.

Fraud detector utilizes machine learning to analyze data, drawing from over 20 years of Amazon's fraud detection expertise.

Users can create customized fraud detection models, interpret the model's evaluations through decision logic, and assign outcomes (like pass or send for review) based on these evaluations.

The best part is it doesn't require users to have machine learning expertise, making it accessible for a wide range of applications.

# Features

01



Uses Machine Learning

02



Fully managed

03



Real-time fraud  
detection

04



Pre-built and  
customizable detectors

05



Built-in Data Encryption  
and Access  
Management

© Copyright KodeKloud

Uses Machine Learning:

Amazon Fraud Detector utilizes machine learning to identify potential fraud activity. The service automatically identifies relevant data, creates and trains a fraud detection model, and deploys it, making it easier to catch fraudulent activity faster than traditional methods.

Fully Managed:

Amazon Fraud Detector is a fully managed service, meaning you don't need to have machine learning expertise to use it. You can create a fraud detection model with just a few clicks in the AWS Management Console, and the service handles all the heavy lifting associated with processing large data sets and training ML models.

#### Real-Time Fraud Detection:

The service can detect potentially fraudulent activity in real-time. By using your data, Amazon Fraud Detector analyses activities as they happen and identifies any actions that may seem out of the ordinary or align with common fraudulent behavior.

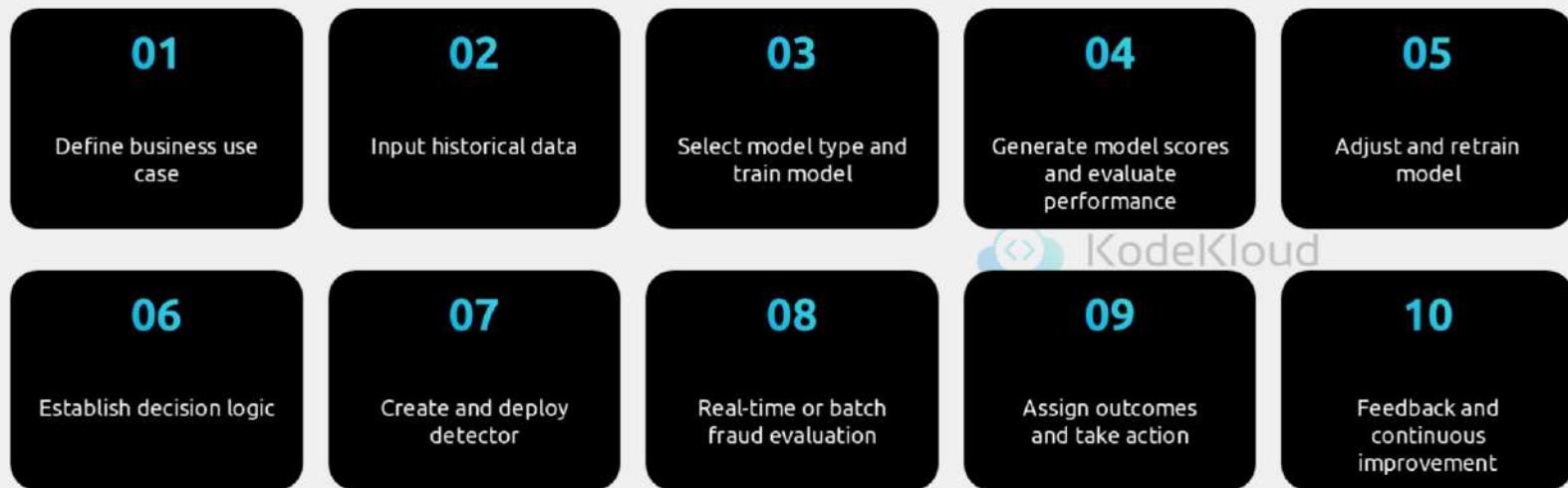
#### Pre-Built and Customizable Detectors:

AWS Fraud Detector provides pre-built templates based on common fraud scenarios (like online payment fraud, account registration fraud, etc.) so you can deploy a model quickly. You can also customize detectors by combining your data, ML models, and rules to suit your specific needs.

#### Built-In Data Encryption and Access Management:

Security is a top priority, and Amazon Fraud Detector includes built-in encryption and robust access management controls. Data is encrypted at rest and in transit, and you can control access to your resources with AWS Identity and Access Management (IAM), ensuring that sensitive information is kept secure.

# Fraud Detector



© Copyright KodeKloud

## Define Business Use Case:

Start with the specific type of fraud you want to detect (e.g., online payment fraud, fake account creation).

## Input Historical Data:

Provide Amazon Fraud Detector with historical data relevant to the defined use case.

### Select Model Type & Train Model:

Based on the use case, Amazon Fraud Detector recommends a model type.  
The system automatically trains the model using the provided data.

### Generate Model Scores & Evaluate Performance:

After training, the model produces scores indicating potential fraud risk.  
Evaluate model performance using these scores and other metrics.

### Adjust & Retrain Model (if necessary):

If model performance is not satisfactory, adjust the input data or settings.  
Retrain the model to improve accuracy.

### Establish Decision Logic (Rules):

Configure rules that dictate how the model should interpret data and assign outcomes (e.g., approve, review, high risk).

### Create & Deploy Detector:

A detector, containing your model and rules, is prepared.  
Test the detector to ensure it operates as expected.  
Deploy the detector to your production environment.

### Real-time or Batch Fraud Evaluation:

The deployed detector analyzes incoming data in real-time or evaluates batched data offline.  
It compares new data with historical patterns and applies the predefined rules.

### Assign Outcomes & Take Action:

Based on the analysis, the detector assigns outcomes to each activity.  
Your system takes action depending on the assigned outcome (e.g., allow, review, block).

### **Feedback & Continuous Improvement:**

Collect feedback on the detector's performance.

Use this feedback and additional data to continuously improve the model and rules.

# How Fraud Detector Works



© Copyright KodeKloud

Refer this site to learn more about: <https://aws.amazon.com/blogs/machine-learning/reviewing-online-fraud-using-amazon-fraud-detector-and-amazon-a2i/>

Online fraud is a significant global issue, costing organizations billions annually.

Machine Learning (ML) is crucial in detecting online fraud, but it requires large labeled datasets, expertise, infrastructure,

and human review systems for high-risk predictions, which are often challenging and costly to implement.

Amazon Fraud Detector is a fully managed service that uses ML and Amazon's extensive experience in fraud detection. It automates the steps to build, train, and deploy an ML model for fraud detection, tailoring each model to your specific dataset, thus improving accuracy over generic solutions.

Amazon A2I makes it easy to build human review workflows for ML models, providing a human-in-the-loop system for any developer.

The solution involves sending information to Amazon Fraud Detector, which assesses the risk score. High-risk predictions are sent for human review through Amazon A2I, and the results are stored in Amazon Simple Storage Service (Amazon S3).

# Fraud Detector



© Copyright KodeKloud

## Fake or Abusive Reviews

The event of interest is the posting of a product review which may contain misleading or abusive content. Automating screening is critical for scaling the ability to spot fake and abusive reviews so that customer service teams don't have to wade through mountains of alerts, many of which may be false positives.

# Amazon Transcribe



# Transcribe



© Copyright KodeKloud

Amazon Transcribe is an AWS Artificial Intelligence (AI) service that makes it easy for you to convert speech to text.

Using Automatic Speech Recognition (ASR) technology, you can use Amazon Transcribe for a variety of business applications, including transcription of voice-based customer service calls, generation of subtitles on audio/video content, and taking precise notes during meetings, enhancing productivity and capturing important conversations accurately.

# Features

01



Automatic speech  
recognition

02



Real-time  
transcription

03



Custom vocabulary

04



Speaker  
identification

05



Integration with  
other AWS services

© Copyright KodeKloud

## Automatic Speech Recognition:

At its core, AWS Transcribe uses advanced deep learning technologies to recognize speech in audio files and transcribe them into text. It supports multiple languages and can handle various accents and dialects effectively.

## Real-Time Transcription:

AWS Transcribe can transcribe audio streams in real time. This feature is particularly useful for applications such as live

subtitling, meeting transcriptions, or real-time closed captioning.

#### Custom Vocabulary:

Users can expand the AWS Transcribe service's recognition ability by adding custom words or phrases specific to their industry, such as product names, technical terminology, or acronyms. This improves the accuracy of transcriptions in specialized or technical contexts.

#### Speaker Identification:

The service can identify when different speakers take turns speaking, even in settings with multiple individuals speaking. This feature, known as speaker diarization, makes it easier to understand who said what in the transcript, which is especially useful for transcribing meetings, interviews, or shows with several participants.

#### Integration with other AWS services:

Amazon Transcribe integrates seamlessly with other AWS services, allowing you to easily process audio files stored in Amazon S3

# Transcribe – Speaker Diarization



© Copyright KodeKloud

With speaker diarization, you can distinguish between different speakers in your transcription output. Amazon Transcribe can differentiate between a maximum of 10 unique speakers and labels the text from each unique speaker with a unique value (spk\_0 through spk\_9).

# Transcribe



© Copyright KodeKloud

**Amazon S3 (Simple Storage Service):**  
Store and retrieve audio files and transcription outputs.

**AWS Lambda:**  
Automatically run code in response to triggers from Amazon Transcribe or Amazon S3.

**Amazon Translate:**

Translate transcribed text into different languages for global accessibility.

**Amazon Comprehend:**

Perform natural language processing (NLP) on transcribed text to extract insights, entities, key phrases, sentiment, and more.

**Amazon DynamoDB:**

Store and retrieve transcription results and metadata for quick access and querying.

# Transcribe



© Copyright KodeKloud

## Customer Conversations

With Transcribe Call Analytics, businesses can extract actionable insights from customer conversations. This feature is beneficial for improving customer engagement, increasing agent productivity, and providing quality management alerts to supervisors.

# Amazon Translate



# Translate



© Copyright KodeKloud

Amazon Translate is a neural machine translation service that offers fast, high-quality, affordable, and customizable language translation.

Amazon Translate offers a free tier with millions of characters free of charge.

It supports 5,550 language combinations.

In this example, chatbot was programmed for English speaking audience. Integrating it with AWS translates can provide

instant support to users in their native language. This breaks the language barrier, offering a more inclusive and global service.

For global businesses or platforms with a diverse user base, this ensures that users can receive immediate assistance in their preferred language, enhancing user satisfaction and engagement.

Instead of employing multilingual support staff or multiple chatbots programmed in different languages, businesses can leverage AWS Translate to efficiently cater to all language requirements. This not only saves resources but also ensures consistent quality of support across languages.

It's particularly useful during high-traffic periods or for 24/7 support, as it can handle unlimited users simultaneously without additional human resources.

# Translate



© Copyright KodeKloud

Source text—The text that you want to translate. You provide the source text in UTF-8 format.

Output text—The text that Amazon Translate has translated into the target language. Output text is also in UTF-8 format. Depending on the source and target languages, there might be more characters in the output text than in the input text.

The translation model has two components, the encoder and the decoder. The encoder reads a source sentence one word at a time and constructs a semantic representation that captures its meaning. The decoder uses the semantic

representation to generate a translation one word at a time in the target language.

# Features

01



Neural Machine Translation

02



Wide range of supported languages

03



Real-time translation

04



Seamless integration

05



Custom terminology

© Copyright KodeKloud

Neural Machine Translation (NMT):

AWS Translate uses neural machine translation techniques to provide more accurate and natural-sounding translation than traditional statistical and rule-based translation algorithms. The service continually learns and improves over time as it ingests more data.

Wide Range of Supported Languages:

Amazon Translate supports translation between a substantial number of languages. This includes widely spoken languages like English, Spanish, Hindi French, German, Chinese, and many more, covering a significant portion of the world's population.

#### Real-Time Translation:

AWS Translate can perform real-time translation, which is crucial for applications that require immediate feedback, such as live chat, online customer support, and real-time multilingual communications.

#### Seamless Integration:

Amazon Translate integrates easily with other AWS services, such as Amazon S3, Amazon Polly, and AWS Lambda, allowing developers to create end-to-end solutions with translation capabilities. For instance, you can translate large volumes of text stored in Amazon S3, or trigger a translation action in response to an event using AWS Lambda.

#### Custom Terminology:

With Amazon Translate, you can create a custom terminology that the service will recognize and apply when translating text. This is crucial for businesses and industries that use specialized jargon or need to maintain brand consistency.

# Translate



© Copyright KodeKloud

**Call Initiation:** A customer initiates a call, and AWS Transcribe starts transcribing the speech into text in real time.

**Language Detection and Translation:**

The system can either automatically detect the language being spoken or the customer can specify their preferred language.

AWS Translate then translates the transcribed text into the desired language in real time. This can work both ways -

translating the customer's speech for the agent and translating the agent's response for the customer.

**Communication and Response:**

For text-based communication (like chat support), the translated text can be presented directly to the agent and customer.

For voice communication, AWS Polly can convert the translated text back into speech, which is then relayed to the customer. This allows for automated, multilingual voice responses based on agent input or scripted responses.

# Translate



© Copyright KodeKloud

Enable multilingual user experiences in your applications by integrating Amazon Translate

In this example, user can choose their native language and use application.

Here source HTML page would be translated into selected language

# Translate



© Copyright KodeKloud

## Amazon Comprehend:

Perform natural language processing (NLP) on texts to extract insights and relationships.  
Use AWS Translate to convert content into preferred languages before or after NLP.

# Amazon Textract



# Textract



© Copyright KodeKloud

Current Challenges: Manual data entry, time-consuming, prone to errors, expensive.

AWS Textract is a service that uses machine learning to automatically extract text and data from scanned documents.

Here, it will be used to digitize and process the information contained in various medical reports. This can include converting handwritten notes to text, extracting information from tables, and identifying key data points. Once the data is

extracted and processed, it will be stored into database where patient records and medical reports are stored

Benefits: Increases efficiency, improves accuracy, enhances data accessibility, enables detailed analysis.

# Document Extraction – Categories

**01**



Text Extraction

**02**



Form Extraction

**03**



Table Extraction

**04**



Signatures

**Text Extraction:** Extracts raw text from a document.

**Form Extraction:** Identifies key-value pairs in forms, linking form data to extracted text.

**Table Extraction:** Extracts tables, cells, items within cells, and other table-related information. Can return results in JSON, CSV, or TXT formats.

Signatures: Detects locations of signatures in documents, returning geometry objects with bounding boxes.

Queries: Allows the addition of queries to specify information needed from a document, returning the information in a separate response structure.

# Features

01



Text extraction

02



Data extraction from  
forms and tables

03



Integration

04



Handwriting  
recognition

05



Scalability

© Copyright KodeKloud

## Text Extraction:

Amazon Textract can accurately extract text in multiple languages from documents, forms, and tables. It goes beyond simple Optical Character Recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

## Data Extraction from Forms and Tables:

Beyond just extracting text, Textract can understand the data stored in forms and tables, recognizing the key-value pairs and tabular data to accurately extract this information for various applications.

#### **Handwriting Recognition:**

Unlike basic OCR tools, Textract can also recognize handwritten text from scanned documents with high accuracy, which is a standout feature.

#### **Integration:**

Amazon Textract easily integrates with other AWS services like Amazon S3, AWS Lambda, and AWS Comprehend. This allows for powerful document processing pipelines.

#### **Scalability:**

It's also highly scalable, processing millions of document pages in a short time, with no machine learning experience required.

# Textract



© Copyright KodeKloud

In this architecture,  
S3 bucket acts as the initial storage location where a text file is uploaded.  
The Lambda function is set up to be triggered automatically when a new file is uploaded to the S3 bucket.  
Once the Lambda function is triggered, it uses Amazon Textract to extract text from the uploaded file.  
Textract processes the uploaded file and extracts the text content from it.  
After the text has been extracted using Textract, the Lambda function then stores the extracted text back into the S3

bucket, either as a new file or by updating the existing file.

# Textract



© Copyright KodeKloud

In this example, when a text file is uploaded to an Amazon S3 bucket, an AWS Lambda function is automatically triggered. This function utilizes AWS Textract to extract textual content from the uploaded file. Once extracted, the text is then stored in an AWS DynamoDB table for further processing and retrieval.

# AWS Glue



# Glue



# Glue

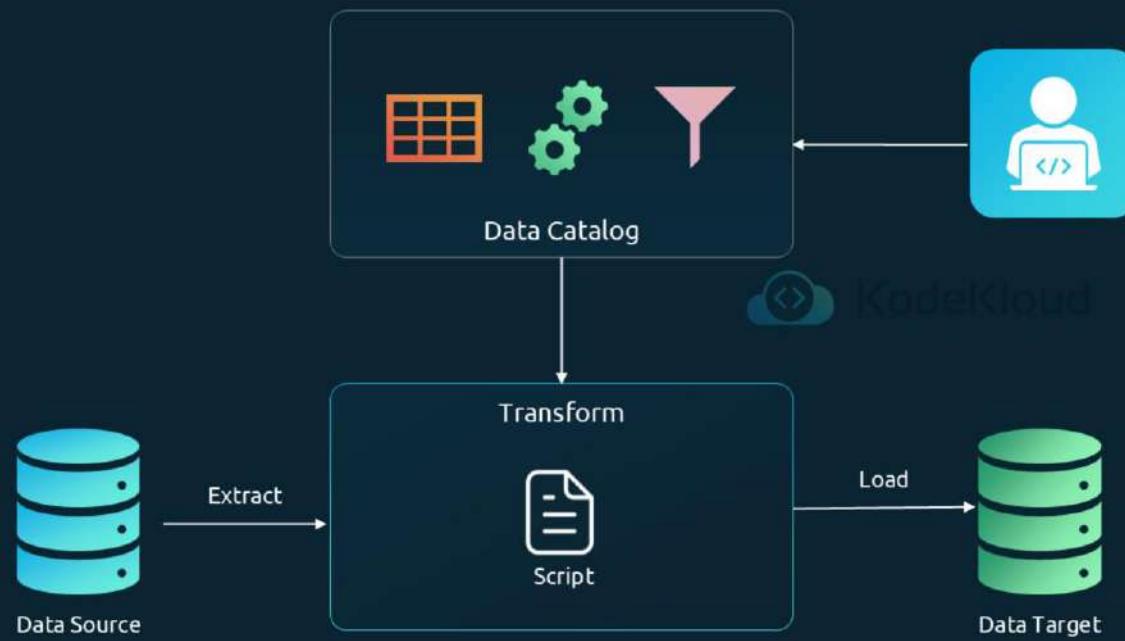


© Copyright KodeKloud

For data store sources, you define a crawler to populate your AWS Glue Data Catalog with metadata table definitions. You point your crawler at a data store, and the crawler creates table definitions in the Data Catalog.

In addition to table definitions, the AWS Glue Data Catalog contains other metadata that is required to define ETL jobs. You use this metadata when you define a job to transform your data.

# Glue



© Copyright KodeKloud

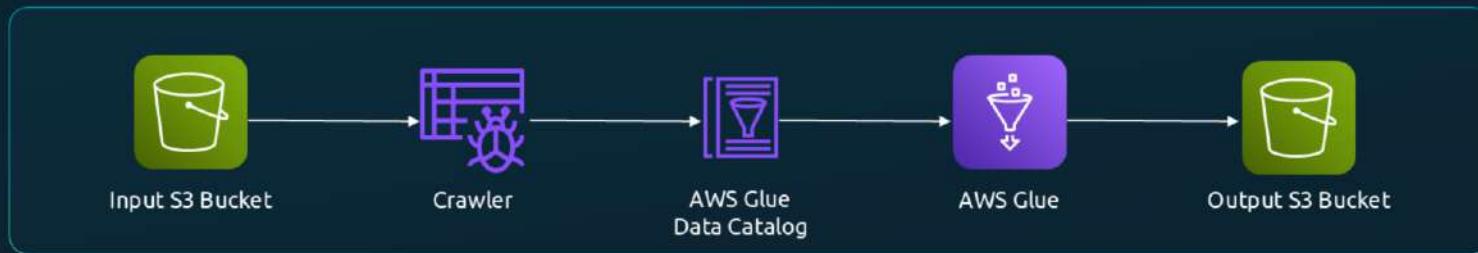
AWS Glue can generate a script to transform your data. Or, you can provide the script in the AWS Glue console or API.

You can run your job on demand, or you can set it up to start when a specified trigger occurs. The trigger can be a time-based schedule or an event.

When your job runs, a script extracts data from your data source, transforms the data, and loads it to your data target. The

script runs in an Apache Spark environment in AWS Glue.

# Glue



© Copyright KodeKloud

## How AWS Glue works?

**Input S3 bucket:** This is the Amazon Simple Storage Service (S3) bucket where your input data is stored.

**Crawler:** An AWS Glue Crawler is used to scan the contents of your input data in the S3 bucket. It automatically identifies the schema of your data (i.e., the structure of your data, such as the column names and data types) and classifies the data format (e.g., JSON, CSV, Parquet).

**AWS Glue Data Catalog:** After the crawler has processed the data, it writes metadata to the AWS Glue Data Catalog, which acts as a central repository to store structural and operational metadata for all your data. It's a managed service that lets you store, annotate, and share metadata in the AWS Cloud.

**AWS Glue:** This represents the ETL jobs that you can define within AWS Glue. These jobs use the metadata in the Data Catalog to carry out the transformation of your data. You can write scripts in PySpark or Scala, or you can use the provided graphical interface to transform your data without writing code. The ETL job will read your data, perform transformations and processing as per your requirements, and then it's ready to be written to the target.

**Output S3 bucket:** Finally, after the transformation, the processed data is written back to another S3 bucket (output S3 bucket). This is where your transformed data resides, and you can use it for further analysis or processing with other AWS services such as Amazon Redshift, Amazon Athena, or Amazon QuickSight.

# Features

01



Serverless ETL service

02



Data catalog

03



Automatic schema discovery

04



Visual ETL job authoring

05



Built-in data transformation libraries

© Copyright KodeKloud

## Serverless ETL Service:

AWS Glue is a serverless service, meaning that you don't need to provision or manage any resources, and you pay only for resources consumed during the execution of your jobs. This serverless approach removes the need for infrastructure management, making it easier for you to focus on developing your ETL jobs.

## Data Catalog:

AWS Glue features a centralized Data Catalog, which is a persistent metadata store for all your data assets, regardless of where they are located. The Data Catalog is automatically populated with metadata from your source data and can be used as a metadata repository for other AWS services like Amazon Athena, Amazon Redshift, and Amazon EMR.

#### Automatic Schema Discovery:

AWS Glue automatically scans data in various data stores, identifies the format, and infers the schema. It then stores this metadata in its Data Catalog, making your data immediately searchable and queryable. This feature simplifies the process of understanding and working with new data formats and sources.

#### Visual ETL Job Authoring:

AWS Glue Studio is an integrated development environment (IDE) that allows you to visually create, run, and monitor ETL jobs in AWS Glue. You can use a visual interface to compose data transformation workflows and visualize the flow of data between different stages of the ETL process.

#### Built-in Data Transformation Libraries:

AWS Glue provides built-in transformation capabilities, called transforms, which include a set of common data transformations like renaming fields, filtering records, joining and aggregating data from different sources, converting data formats, or cleaning and normalizing data and converting data types. These built-in functions mean you don't have to write the code from scratch, simplifying the ETL job creation process.

# Glue



© Copyright KodeKloud

## AWS data pipeline

Data Sources: Amazon S3, Amazon Redshift, Amazon RDS, DB running on EC2

Crawler: An AWS Glue Crawler is used to inspect the different data sources. It will automatically discover and profile the data, and then create metadata tables in the AWS Glue Data Catalog.

AWS Glue Data Catalog: The metadata tables created by the crawler are stored in the Data Catalog, which provides a

unified metadata repository for all your data assets.

AWS Glue ETL: This is where the actual ETL process happens. AWS Glue ETL jobs use the metadata stored in the Data Catalog to understand the data's schema. The ETL job then reads, transforms, and processes the data accordingly.

Data Targets/Analytics Tools: AWS services that can be used for further data processing, analytics, and business intelligence after the ETL process.

Amazon Redshift: Post-ETL, the transformed data can be loaded into Redshift for complex analytics and business intelligence workloads.

Amazon Athena: An interactive query service that allows you to analyze data in Amazon S3 using standard SQL.

Amazon EMR: Amazon Elastic MapReduce can be used for processing vast amounts of data using open-source tools such as Spark, Hadoop, etc.

QuickSight: An analytics service that you can use to visualize the data and perform business intelligence.

# AWS Glue DataBrew



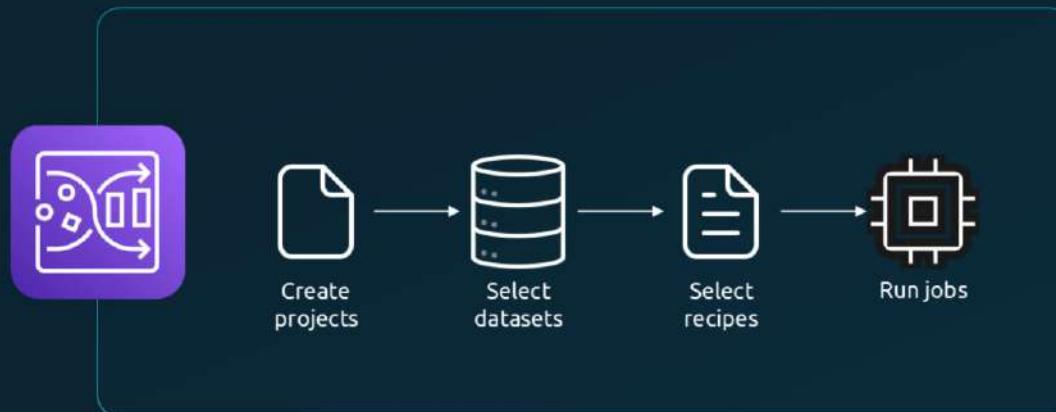
# Glue DataBrew



© Copyright KodeKloud

AWS Glue DataBrew, which is a visual data preparation tool that enables users to clean and normalize data without writing code.

# Glue DataBrew



© Copyright KodeKloud

AWS Glue DataBrew, which is a visual data preparation tool that enables users to clean and normalize data without writing code.

**Create Projects:** The initial step in the DataBrew workflow where you create a project. A project in DataBrew is a workspace where you can interactively explore, analyze, and perform data preparation tasks on datasets.

Select Datasets: After creating a project, the next step is to select datasets. In AWS Glue DataBrew, datasets are the collections of data you want to analyze and transform. You can import datasets from various sources such as Amazon S3, Amazon Redshift, or other AWS services.

Select Recipes: Once a dataset is chosen, you create or select recipes. Recipes are a set of data transformation steps that you can apply to your datasets. These can include operations like filtering rows, converting data types, and more complex transformations. In DataBrew, recipes are created visually by applying transformations from a palette of operations.

Run Jobs: The final step is to run jobs. A job is the execution of the recipe you've defined on the selected dataset. When you run a job, DataBrew will apply all the transformations specified in the recipe to the entire dataset. The results are then typically stored in Amazon S3, and can be used for analysis or for machine learning applications.

This process allows you to prepare your data for analysis or machine learning in a scalable, serverless environment provided by AWS.

# Glue DataBrew



© Copyright KodeKloud

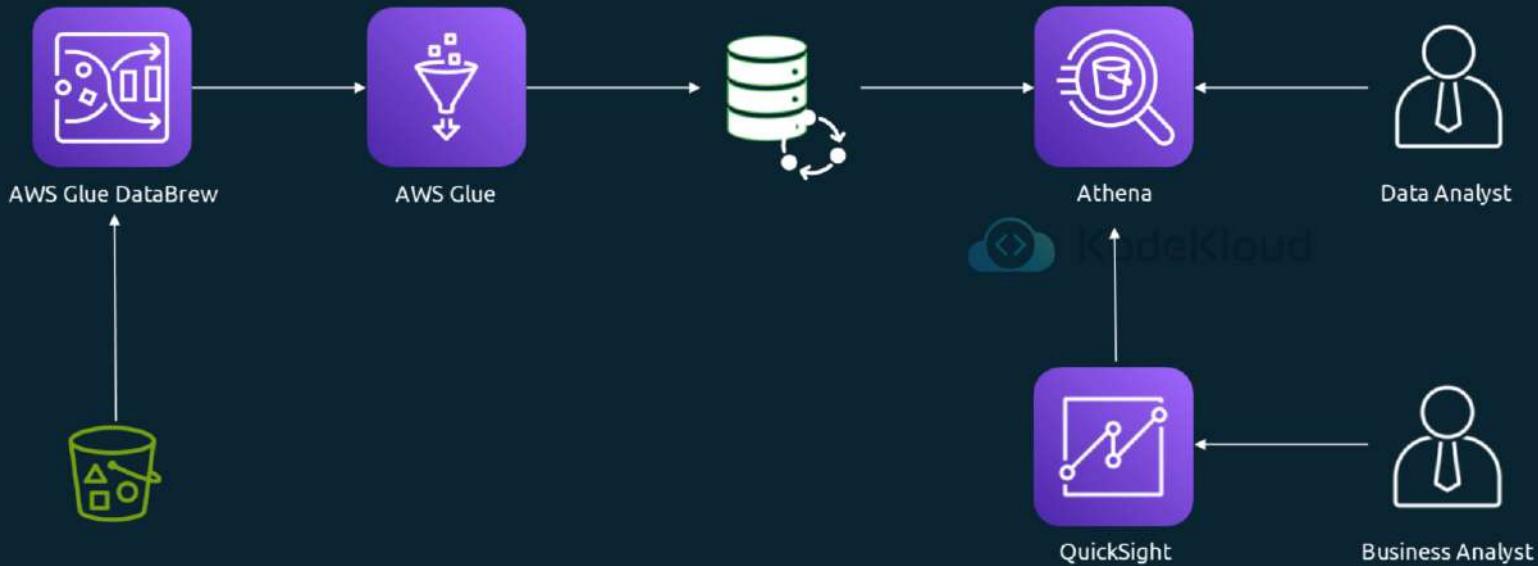
## Data Ingestion:

**Amazon S3:** DataBrew can access files stored in Amazon S3 buckets, which is a common repository for raw data in various formats like CSV, JSON, Parquet, etc.

**AWS Glue Data Catalog:** The Data Catalog provides a persistent metadata store for DataBrew, allowing users to define and manage datasets based on the metadata.

**Database Services:** DataBrew can connect to AWS database services such as Amazon RDS and Amazon Redshift, as well as third-party databases, to import data directly.

# Glue DataBrew



© Copyright KodeKloud

## Data Preparation with AWS Glue Databrew:

AWS Glue Databrew allows users to clean and normalize data without writing code.

It provides over 250 pre-built transformations for data preparation.

Users can visualize their data and the impact of transformations in real-time.

Databrew can handle data from various sources like data lakes, data warehouses, and SaaS applications.

## Data Transformation with AWS Glue ETL:

AWS Glue ETL (Extract, Transform, Load) is used to transform raw data into valuable insights.

It's serverless and can be scaled as needed without user intervention.

Glue ETL has a visual interface, and users can also write custom scripts in Python or Scala.

It integrates with various AWS services and can read data from and write data to different data stores.

# Glue DataBrew



© Copyright KodeKloud

DataBrew can prepare data for machine learning workflows

# Features

01



Visual data preparation

02



Data profiling

03



Scalability and performance

04



Integration with AWS Data Stores

05



Job scheduling and reusability

© Copyright KodeKloud

## Visual Data Preparation:

DataBrew offers a visual interface that allows users to easily apply transformations to their data without writing any code. It provides over 250 pre-built transformations for tasks such as filtering anomalies, standardizing formats, and correcting invalid values, making data preparation faster and more intuitive.

## Data Profiling:

Before the transformation process, DataBrew can automatically profile your data, providing statistics that help you understand data quality and structure. This feature helps identify data inconsistencies, missing values, and anomalies, providing insights that guide the data preparation process.

#### Scalability and Performance:

AWS Glue DataBrew can handle datasets of any size. Whether you're working with a few records or a dataset with billions of records, DataBrew scales to meet your needs, processing data quickly and efficiently.

#### Integration with AWS Data Stores:

DataBrew integrates seamlessly with various AWS data stores such as Amazon S3, Amazon Redshift, Amazon Aurora, and others. This means you can easily import data from and export data to these services, streamlining the data preparation process within the AWS ecosystem.

#### Job Scheduling and Reusability:

You can automate data preparation tasks by scheduling recurring DataBrew jobs. This is particularly useful for datasets that require regular updates. Additionally, you can create project templates that can be reused across different datasets, making consistent data preparation processes across different projects or teams.

# Amazon Elastic MapReduce (EMR)



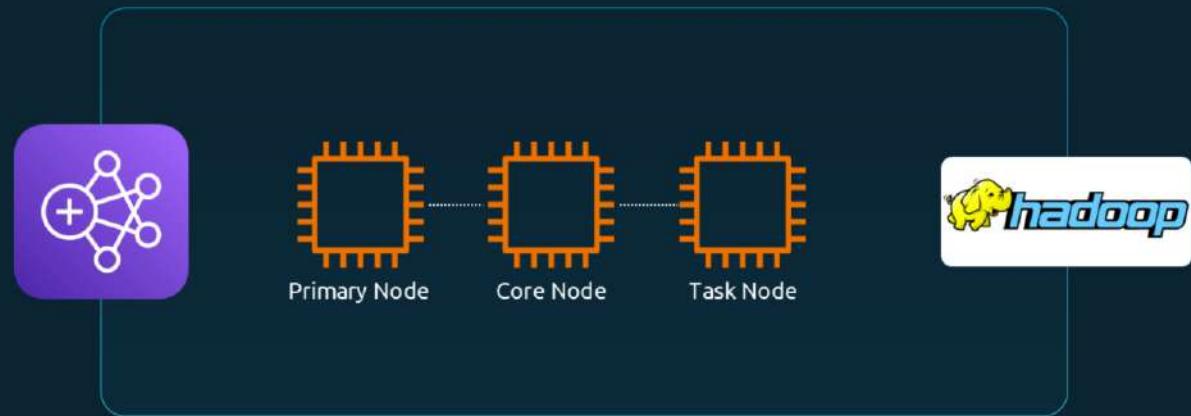
# Elastic MapReduce (EMR)



© Copyright KodeKloud

Amazon EMR (previously called Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. Using these frameworks and related open-source projects, you can process data for analytics purposes and business intelligence workloads. Amazon EMR also lets you transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

# Elastic MapReduce (EMR)



© Copyright KodeKloud

The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node. Each node has a role within the cluster, referred to as the node type. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

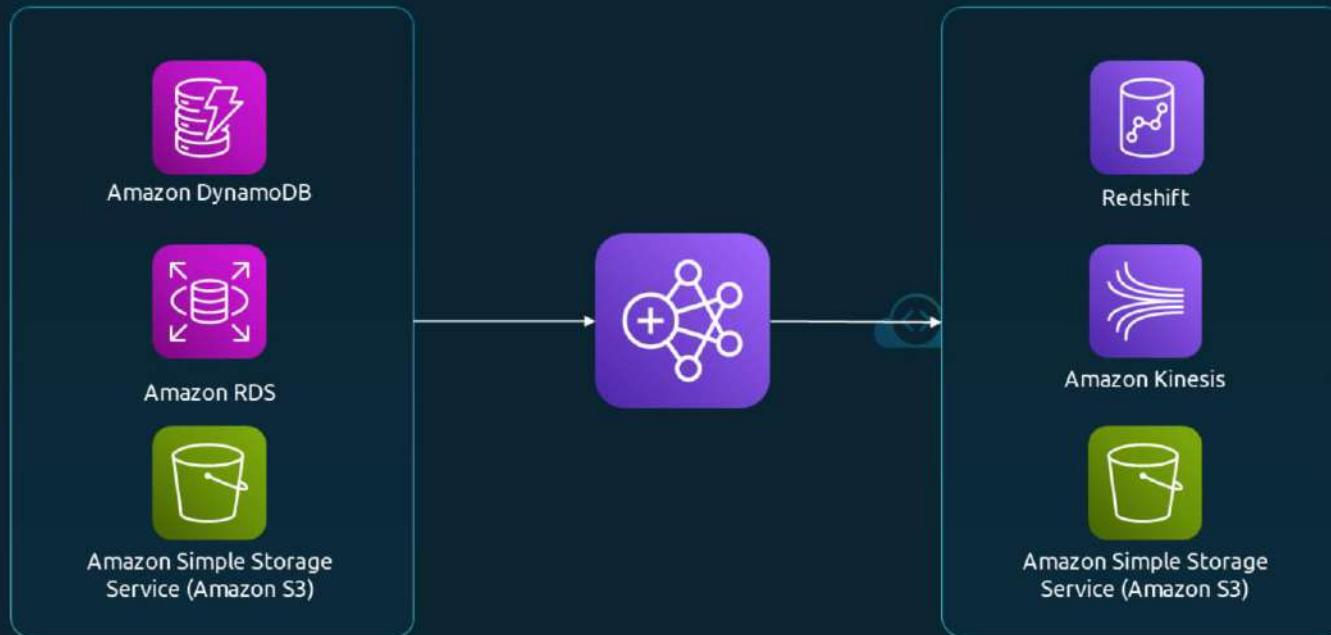
The node types in Amazon EMR are as follows:

**Primary node:** A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes for processing. The primary node tracks the status of tasks and monitors the health of the cluster. Every cluster has a primary node, and it's possible to create a single-node cluster with only the primary node.

**Core node:** A node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node.

**Task node:** A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional.

# Elastic MapReduce (EMR)

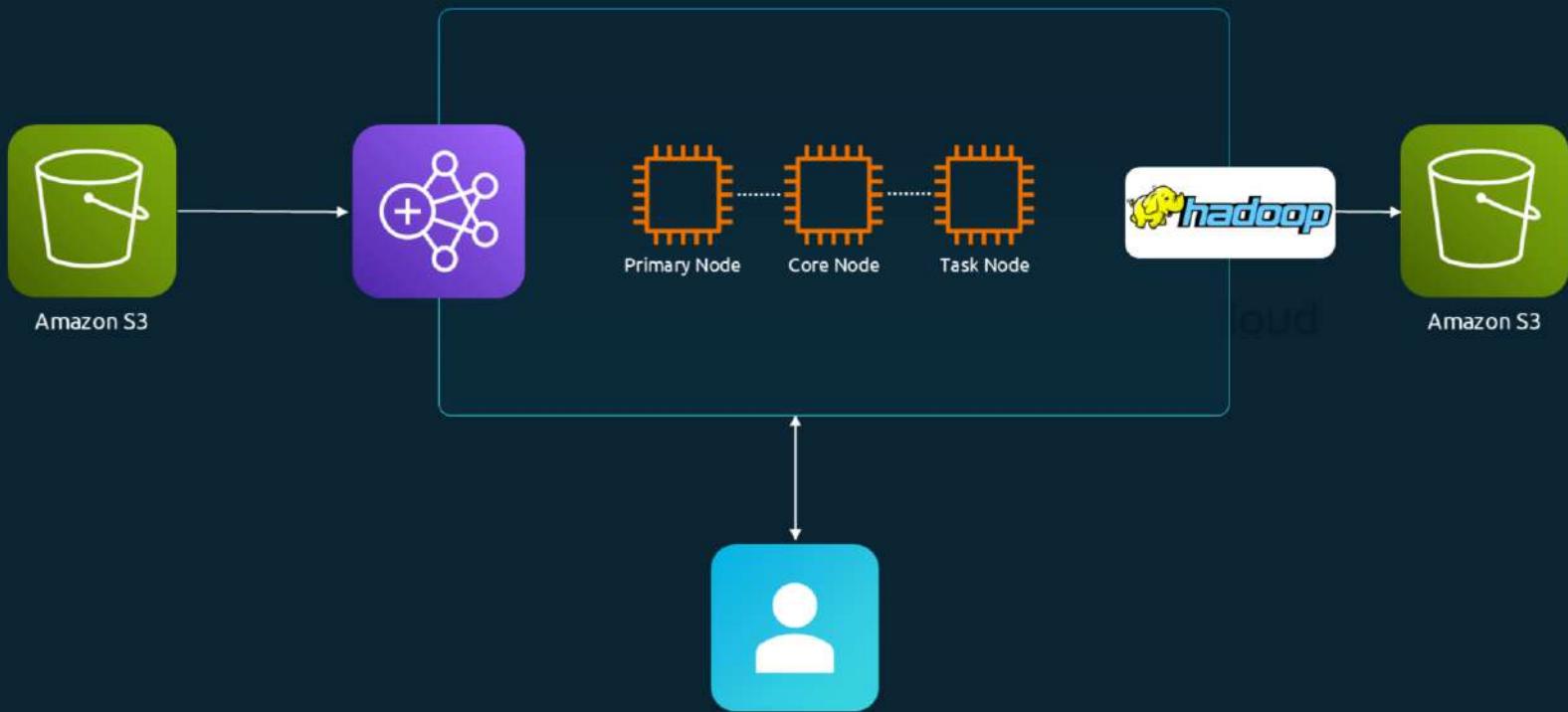


© Copyright KodeKloud

AWS EMR is a cloud service that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data.

EMR can pull large datasets from services (Amazon DynamoDB, Amazon RDS, Amazon S3), process it using big data frameworks, and output the results to other services (Amazon Redshift, Amazon Kinesis, Amazon S3) for further analysis or storage.

# Elastic MapReduce (EMR)



© Copyright KodeKloud

## How EMR works?

**Launch:** You start by launching an EMR cluster, specifying the number and type of EC2 instances you need based on the size and complexity of your data processing tasks.

**Load Data:** Data is loaded into the EMR file system from various sources. This could be from Amazon S3 (Simple Storage Service), DynamoDB, or any other supported AWS data storage service.

**Process Data:** EMR uses Hadoop to distribute your data across the cluster and process it in parallel. You can run various big

data frameworks and processing engines such as Apache Spark, HBase, Presto, and Hive on your cluster. Monitor and Manage: AWS provides tools like the EMR console, the AWS CLI (Command Line Interface), and the EMR API to monitor the cluster's performance, scale the number of instances up or down, and manage the cluster's lifecycle.

Output Data: After processing, the output data can be written back to S3, passed to another AWS service for further processing, or used to generate reports.

Terminate Cluster: Once the processing is complete, you can terminate the cluster to stop incurring charges. Since EMR is elastic, you only pay for the resources you use.

# Elastic MapReduce (EMR)



© Copyright KodeKloud

## Running steps to process data

You can submit one or more ordered steps to an Amazon EMR cluster. Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

The following is an example process using four steps:

Submit an input dataset for processing.

Process the output of the first step by using a Pig program.

Process a second input dataset by using a Hive program.

Write an output dataset.

Steps are run in the following sequence:

A request is submitted to begin processing steps.

The state of all steps is set to PENDING.

When the first step in the sequence starts, its state changes to RUNNING. The other steps remain in the PENDING state.

After the first step completes, its state changes to COMPLETED.

The next step in the sequence starts, and its state changes to RUNNING. When it completes, its state changes to COMPLETED.

This pattern repeats for each step until they all complete and processing ends.

If a step fails during processing, its state changes to FAILED. You can determine what happens next for each step. By default, any remaining steps in the sequence are set to CANCELLED and do not run if a preceding step fails. You can also choose to ignore the failure and allow remaining steps to proceed, or to terminate the cluster immediately.

# Features

01



Managed Hadoop framework

02



Scalability and flexibility

03



Cost-effective processing

04



Integration with other AWS services

05



Security and compliance

© Copyright KodeKloud

## Managed Hadoop Framework:

Amazon EMR simplifies big data processing, providing a managed Hadoop framework that distributes data computation across Amazon EC2 instances for processing. EMR supports several popular distributed frameworks such as Apache Hadoop, Apache Spark, and HBase, which are pre-installed and configured in your cluster.

## Scalability and Flexibility:

EMR clusters can be easily resized, and you have the option to optimize both the number and type of instances to fit your specific workload. This means you can start with as little as a single instance and scale to thousands, benefiting from the elasticity of the AWS cloud.

#### Cost-Effective Processing:

With Amazon EMR, you can take advantage of the spot pricing of Amazon EC2 instances, allowing you to bid for spare Amazon EC2 capacity at a potentially lower price. Additionally, EMR pricing is simple and predictable: You pay a per-second rate for every second used, with a one-minute minimum charge.

#### Integration with Other AWS Services:

EMR seamlessly integrates with other AWS services. For example, you can read and write data from and to Amazon S3, use Amazon RDS or Amazon DynamoDB as your transactional data store, use Amazon CloudWatch to monitor your clusters, and use AWS CloudFormation for deployment, among others.

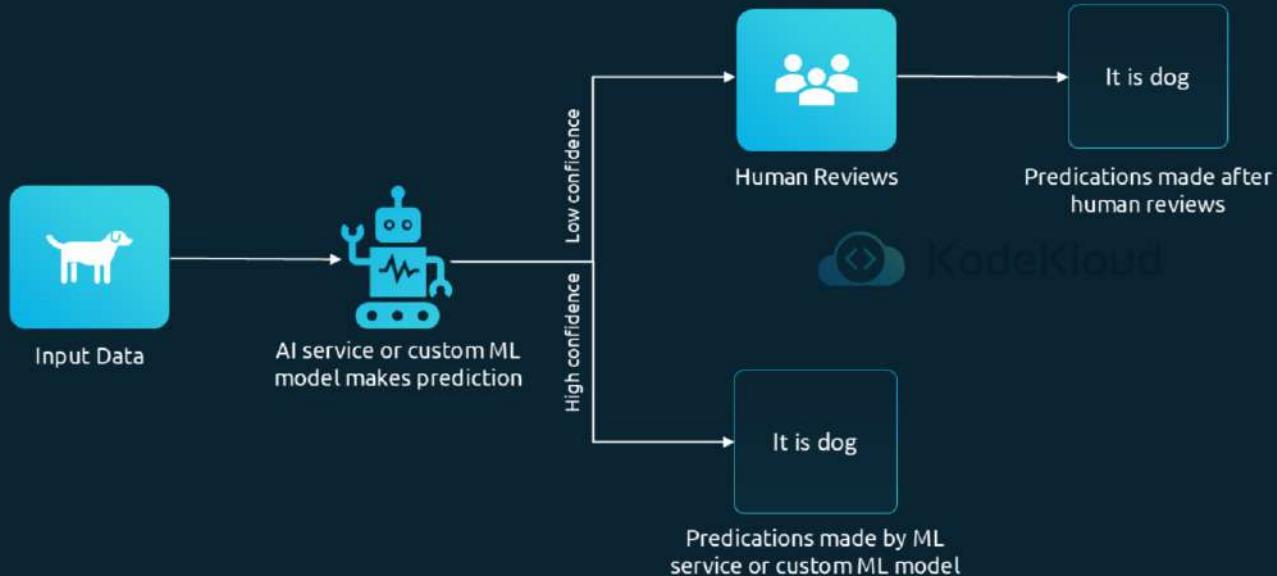
#### Security and Compliance:

Amazon EMR provides multiple layers of security for your data, including AWS Identity and Access Management (IAM), AWS Key Management Service (KMS), encryption at rest and in transit, network isolation with Amazon VPC, and logging with AWS CloudTrail. It also supports compliance requirements like GDPR, HIPAA, and more.

# **Amazon Augmented AI (Amazon A2I)**



# Augmented AI



© Copyright KodeKloud

Amazon Augmented AI (A2I) is an AWS service designed to help developers easily integrate human reviewers into machine learning workflows to review and validate model predictions. While machine learning models can be highly accurate, there are certain use cases where human judgment is essential to verify the model's predictions, especially when accuracy is critical or the problem is complex.

# Features

01



Easy integration  
with ML services

02



Built-in Human  
Review workflows

03



Access to Human  
Reviewers

04



Continuous learning  
and improvement

05



Seamless scaling

© Copyright KodeKloud

## Easy Integration with ML Services:

Amazon A2I can be easily integrated with machine learning services such as Amazon Textract, Amazon Rekognition, and Amazon SageMaker, enabling you to implement human review of predictions efficiently.

## Built-in Human Review Workflows:

It provides the necessary resources to set up human review workflows for common machine learning use cases, such as

content moderation and text extraction from documents, without requiring any code.

**Access to Human Reviewers:**

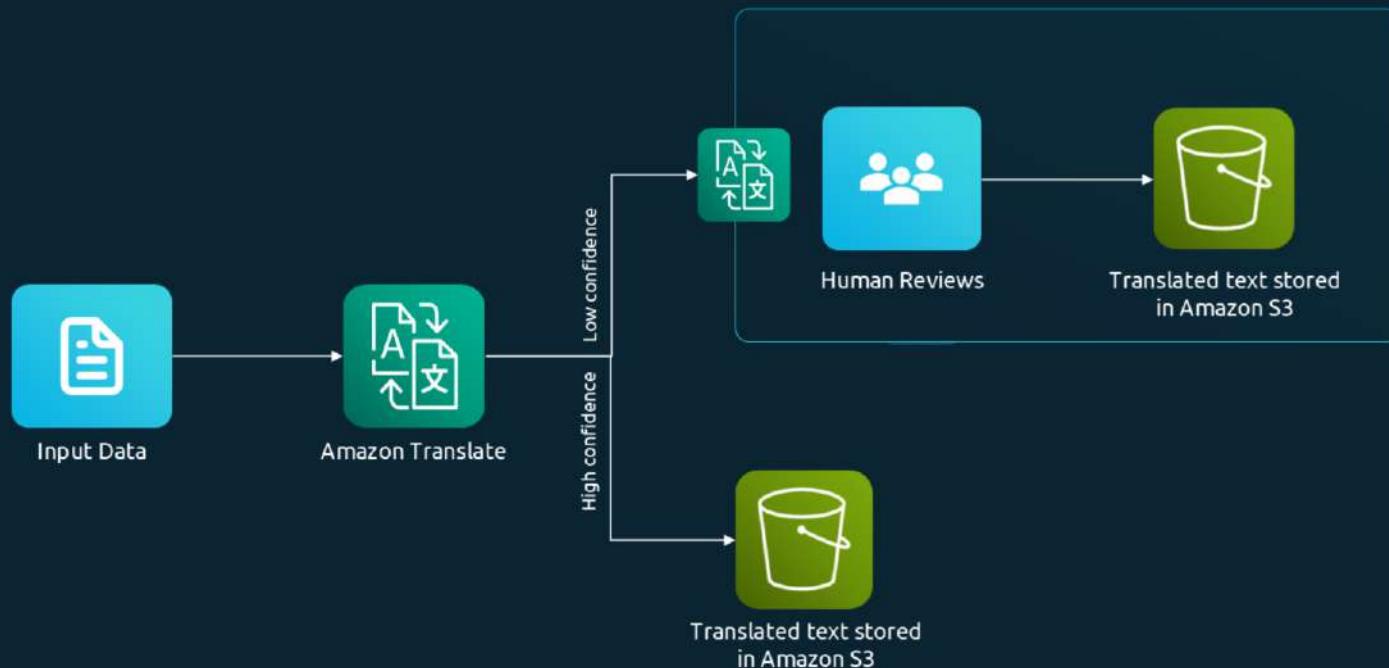
You can either use your own human reviewers or access the workforce of third-party vendors via Amazon Mechanical Turk, a marketplace for work that requires human intelligence.

**Continuous Learning and Improvement:**

By incorporating human decisions back into your machine learning models, you can continuously improve the accuracy of your models over time.

**Seamless Scaling:** The service scales the number of human reviewers based on the volume of predictions to be reviewed, ensuring that you can handle varying workloads without manual intervention.

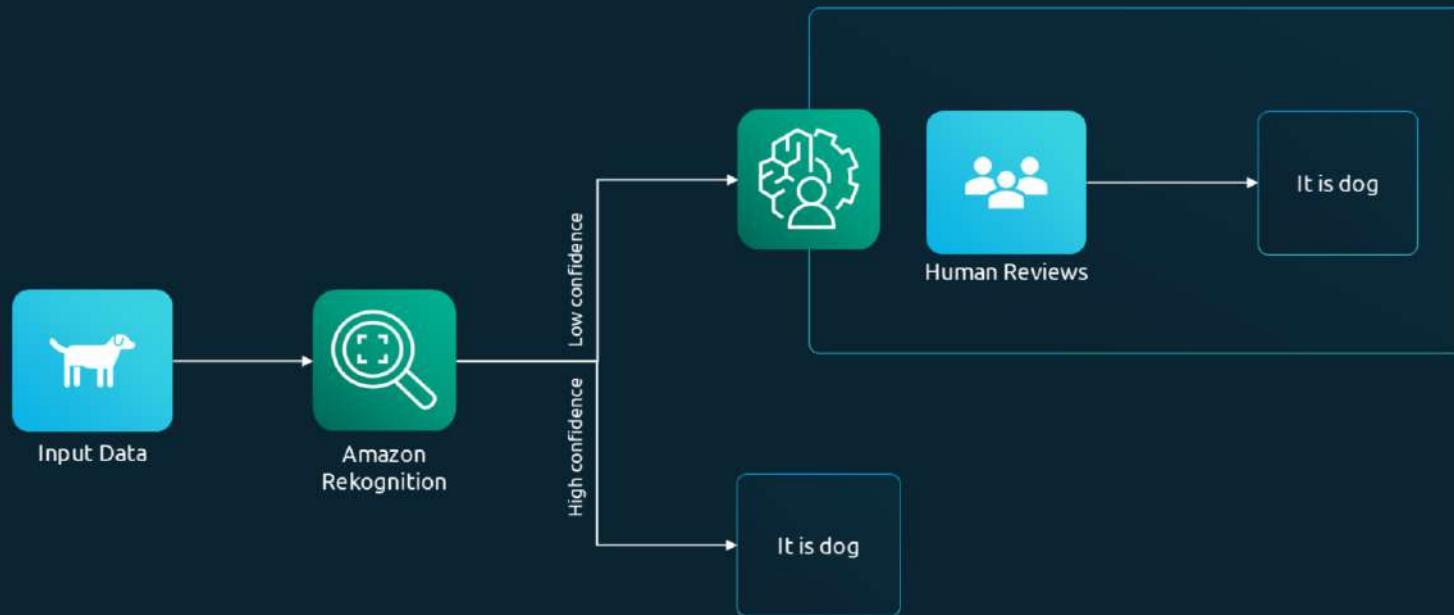
# Augmented AI



© Copyright KodeKloud

For applications using Amazon Translate, A2I can help in reviewing and improving translation outputs, especially in contexts where nuanced translation is critical.

# Augmented AI



© Copyright KodeKloud

Integration with Amazon Rekognition allows you to review image and video analysis results. For instance, you can use A2I to verify Rekognition's content moderation predictions, ensuring inappropriate content is accurately identified and handled.

Similarly, for tasks like facial analysis or object detection where high precision is required, A2I can facilitate human review to confirm or correct the predictions

# Augmented AI



© Copyright KodeKloud

A2I can directly integrate with Amazon Mechanical Turk, providing access to a diverse and scalable workforce. This is especially useful for businesses without their own review teams.

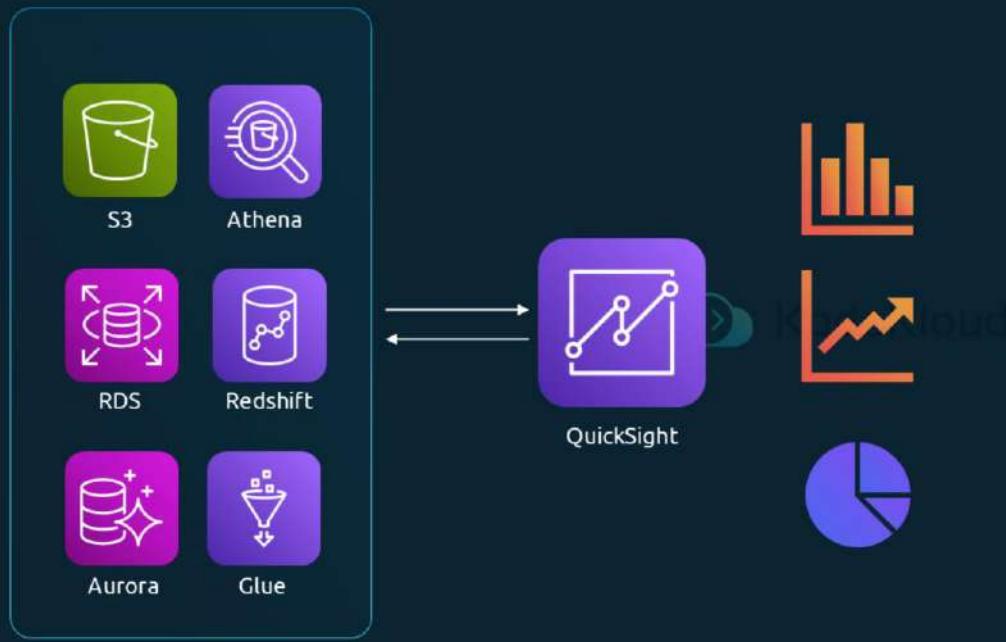
# Amazon QuickSight



# QuickSight



# QuickSight



© Copyright KodeKloud

## Amazon S3:

Directly connect to data stored in S3 buckets.

Use Athena to query data in S3 and visualize the results in QuickSight.

## Amazon Athena:

Direct integration to run ad-hoc queries on data in S3.

Visualize Athena query results in QuickSight dashboards.

Amazon RDS:

Connect to relational databases like RDS MySQL, PostgreSQL, MariaDB, SQL Server, and Oracle.  
Directly import data into SPICE or query the databases in real-time.

Amazon Redshift:

Native integration allows for direct querying of Redshift clusters.  
Leverage Redshift's performance to run complex queries and visualize in QuickSight.

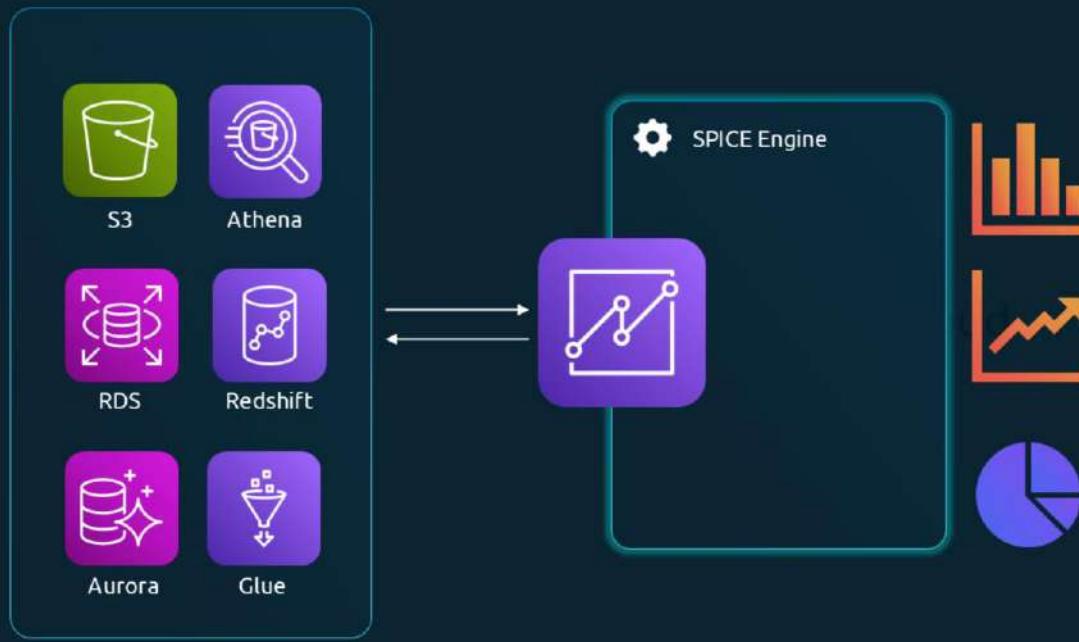
AWS Glue:

Use Glue Data Catalog as a centralized metadata repository.  
Glue ETL jobs can transform data and make it available for QuickSight.

Amazon Aurora:

Connect to both Aurora MySQL and Aurora PostgreSQL.  
Directly visualize data from Aurora in QuickSight.

# QuickSight



© Copyright KodeKloud

In AWS QuickSight, SPICE (Super-fast, Parallel, In-memory Calculation Engine) is the in-memory calculation engine that powers the service's data processing and analysis. It's designed to perform advanced calculations and serve data quickly through the QuickSight dashboards, enabling users to visualize and interact with their data at high speeds.

# SPICE

## Super-fast, Parallel, In-memory Calculation Engine

SPICE provides QuickSight with high-speed data processing for responsive dashboards.



SPICE stores large datasets in-memory, easing the load on databases and reducing query costs.



SPICE datasets can refresh automatically to ensure data on dashboards remains up to date.



SPICE engine scales seamlessly with data volume and complexity, requiring no manual scaling efforts.



Encryption within SPICE ensures that data is secure while at rest.



SPICE supports QuickSight Q for natural language querying and ML-driven insights.



In AWS QuickSight, SPICE (Super-fast, Parallel, In-memory Calculation Engine) is the in-memory calculation engine that powers the service's data processing and analysis. It's designed to perform advanced calculations and serve data quickly through the QuickSight dashboards, enabling users to visualize and interact with their data at high speeds.

# Features

01



Serverless and fully managed

02



SPICE engine

03



Interactive dashboards

04



Data preparation and ML insight

05



Integration with AWS Ecosystem

© Copyright KodeKloud

**Serverless & Fully Managed:** QuickSight is serverless, meaning there's no infrastructure to manage, and it automatically scales with your usage.

**SPICE Engine:** Stands for Super-fast, Parallel, In-memory Calculation Engine. It's designed to rapidly perform advanced calculations and render visualizations.

**Interactive Dashboards:** Users can create interactive dashboards with drill-downs, filters, and insights, which can be accessed from desktop and mobile devices.

**Data Preparation and ML Insight:** QuickSight provides tools for data preparation, including data cleansing, joining, and transformation. It integrates machine learning to provide forecasts, anomaly detection, and natural language narratives.

**Integration with AWS Ecosystem:** QuickSight can directly connect to various AWS data sources such as Amazon RDS, Redshift, Athena, and S3. It also supports non-AWS sources like SQL Server, MySQL, and more.



# KodeKloud

© Copyright KodeKloud

Visit [www.kodekloud.com](http://www.kodekloud.com) to learn more.