What is bias?
Bias is the amount that a model's prediction differs from the target value, compared to the training data. (or)
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Statistical bias:
Statistical bias is defined as the difference between the parameter to be estimated and the mathematical expectation of the estimator.

What is variance?
Variance describes how much random variable differs from its expected value.
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

What is the trade-off between bias and variance?
If our model has very few parameters then it may have low variance and high bias. On the other hand, if our model has large number of parameters then it has high variance low bias. So, we need to find the right/good balance without overfitting and underfitting the data.

Gradient: The gradient is nothing but a derivative of loss function with respect to the weights. It is used to updates the weights to minimize the loss function during the back propagation in neural Network.
Vanishing Gradients: Vanishing gradient occurs when the derivative or slope will get smaller and smaller as we go backward with every layer during backpropagation. When weights update is very small, the training time takes too much longer. A vanishing gradient problem occurs with the sigmoid and tanh activation function because the derivative of the sigmoid and tanh activation functions are between 0 to 0.25 and 0-1. This problem happens because of activation function, not because of the wights.
Exploding gradient: occurs when the derivatives or slopes will get larger and larger as we go backward with every layer during the back propagation. This problem happens because of weights, not because of the activation function.

What is gradient descent
Gradient descent is an optimization algorithm used to find the values of coefficients of a function (f) that minimizes a cost function.
Gradient descent based algorithms are linear regression, logistic regression, neural network etc..

How to make a time-series stationarity?
Differencing the series
Take log of the series
Take the n^{th} root of the series
Combination of all above

What are the forecasting algorithms?
ARIMA (Auto Regressive Integrated Moving Average)
SARIMA: Seasonal ARIMA

ARIMA characterized by 3 terms (p, d, q)
Where p is the order of AR term (Auto regressive)
Q is the order of MA term (Moving Average)
D is the differencing required to make the time series stationary.

ARIMA means it is a Linear Regression model that uses its own lags as predictors.

What is Dickey-Fuller Test:
This test is used for the time series is stationary or not.
The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary.

Null Hypothesis (H0): If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has a time dependent structure.

Null Hypothesis (H1): The null hypothesis is rejected, it suggests the time series does not have a unit root, meaning it is stationary. It does not have time dependent structure.

Example: We interpret the result using p-value from the test. A p-value below the threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

Ex: p-value > 0.05: Fail to reject the null hypothesis (H0), so the data has a unit root and is non-stationary.

p-value <=0.05: reject the null hypothesis (H0), the data does not have a unit root and is stationary.

Stationarity and Non-stationarity
Stationary is a property of a time-series.
Stationary data has not a time dependent structure.
A stationary time-series without seasonal effect.
The statistical values, mean and variance is constant over time.

Stationary data contains:
Self correlated
No global trend
No periodicity
No seasonality

Time series can be split into:
Base level + Trend + seasonality + Error

Time series forecasting with Prophet

The Prophet is open-source library for making forecasts for **univariate** time series datasets. It is developed by Facebook.
A piecewise linear or logistic growth curve trend. Prophet automatically detects changes in trends by selecting change points from the data.
Prophet Forecasting Model
$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$
g(t) = piecewise linear or logistic curve for modeling non-periodic changes in time series.
S(t) = periodic changes (ex. Weekly/yearly seasonality)
H(t) = effects of holidays (user provided) with irregular schedules
$\varepsilon(t)$ = error term

FBProphet: is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, daily seasonality, plus holiday effects.

Why is ReLU better and more often used than Sigmoid in Neural Networks?
The ReLU function is f(x) = max(0,x). Usually this is applied element-wise to the output of some other function, such as a matrix vector product.
One way ReLUs improve the neural networks by speeding up training dataset.
Also, the computational step of ReLU is easy: any negative elements are set as 0.0.
In practice, neural networks with ReLU tend to show better convergence performance than sigmoid.

Activation functions:
Generally, the artificial neurons calculate a weighted sum of its input, adds a bias ($Y= \sum(\text{weight} * \text{input}) + \text{bias}$) and then decide purpose we will use activation function.
Activation functions are:
Step function: just checking the threshold value, if it is greater threshold then y is 1 otherwise, 0.
Linear function (A=CX): A straight line function where activation is proportional to input.
Sigmoid function ($A= 1/(1+e^{-x})$): It is nonlinear nature (not straight line). The output of the function is always going to in range (0,1). The Sigmoid is popular in classification problem.

Tanh function (tanh(x) = 2 sigmoid(2x) -1): similar characteristics of sigmoid function, and it is nonlinear in nature.

Softmax is a classifier at the end of the neural network. That is logistic regression to normalize outputs to values between 0 and 1.

Linear activation functions: Sigmoid, softmax, and tanh
Non-linear activation functions: soft sign, Relu (rectified linear unit), and Leaky Relu

What is data normalization (scaling) and why do we need it?
Normalization is a technique often applied as part of data preparation for Machine Learning.
Data normalization means transforming all variables in the data to a specific range or common scale.
Advantages are: Increased consistency, easier object-to-data mapping.
Normalizing your data is reducing the number of duplicates in your database.
Normalization is a scaling technique, the values rescaled to 0 and 1. It is also known as min-max scaling.
$X'_{normal} = (X-X_{min})/(X_{max}-X_{min})$

Standardization technique:
Is another scaling technique where the values are centered around the mean with unit standard deviation
$X' = (X- \mu)/\sigma$
Normalization and standardization techniques commonly used feature scaling technique.

Cosine similarity: is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0 deg is 1, and it is less than 1 for any angle in the interval (0,infinity] radians. It is thus a judgement of orientation and not magnitude. (see Euclidean_cosine_similarity.ipynb)
(or)
Cosine similarity between two vectors corresponds to their dot product divided by the product of their magnitudes.
$$cos(\theta) = \frac{x \cdot y}{||x|| \cdot ||y||}$$
Euclidean distance is the distance between two vectors (Euclidean_cosine_similarity.ipynb)
$$||x-y||_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

EDA (Exploratory Data Analysis): To analyze, investigate datasets and summarize their main characteristics, often employing data visualization methods etc.

How do you handle missing or corrupted data in a dataset?
If you find missing/corrupted data in a dataset and whether drop those rows or columns, or decide replace with mean or median value.
Calculate mean or median of the data and replace it with the missing values.

The clustering algorithms Data Scientist:
Clustering is a ML technique involving the grouping of data points. Using clustering algorithms to classify these into specific groups.

It is an unsupervised learning method and a famous technique for statistical data analysis.
The significance of clustering algorithms allows to segregate data and group them into different clusters depending on their similarities.
K-means clustering algorithm, Mean-Shift Clustering Algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), Agglomerative Hierarchical Clustering, and Latent Dirichlet Allocation (LDA)

K-means clustering
Is an unsupervised machine learning algorithm that aims to group the observations in a given dataset into clusters.
The number of clusters is provided an input.
It forms the clusters by minimizing the sum of the distance of point from their respective cluster centroid.

How do you know which Machine Learning model you should use?
Generally, categorize the depending on the datasets.
If you labelled data, it is supervised learning problem
If you have unlabeled data and want to find the structure, it is an unsupervised learning problem.
If the output of your model is a number, it is a regression problem
If the output of your model is a class, it is a classifier problem
If the output of your model is a set of input groups, it is a clustering problem.
If you want to detect anomaly, that is anomaly detection

How is KNN different from k-means clustering?
K-Nearest Neighbors is a supervised classification algorithm.
K-Nearest Neighbors work, you need labeled data you want to classify the unlabeled point.
K-means clustering is an un supervised algorithm.
K-means clustering requires un labeled points and cluster them into groups.

Explain how a Receiver Operating Characteristics (ROC) curve works?
The ROC curve is a graphical representation of the contrast between the true positive rates and the false positive rates with different thresholds.
ROC curve is a performance measurement for classification problem.
It tells how much model is capable of distinguishing between classes.

The area under the ROC curve is AUC (Area Under the Curve).
The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

What is Confusion Matrix
It is performance measurement for Machine Learning classification problem where output can be two or more classes.

TP: True Positive: FN: False Negative; FP: False Positive; TN: True Negative
Recall: TP/(TP+FN)
Precision: TP/(TP+FP)
F-measure (F1-score) = 2*Recall*Precision/(Recall + Precision)

What is the F1 score? How would you use it?
F1 score is a measure of model performance.
It is weighted average of the precision and recall of a model.
The result varies between 0 and 1, 1 is best and 0 being the worst.
You would use it in classification models.

Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)

Sensitivity and Specificity measures are used to plot the ROC curve. The area under the ROC curve is AUC (Area Under the Curve).

Explain the difference between L1 and L2 regularization
L1 regularization is called **Lasso** regression, and L2 is called **Ridge** regression.
L1 regularization tries to estimate the median of the data while the L2 regularization tries to estimate the mean of the data to avoid overfitting.
L1 is just sum of the weights, and L2 is the sum of the squares of the weights.
L1 regularization helps in feature selection by eliminating the features that are not important. To avoid overfitting also.

What is the difference between Type I and Type II error?
Type I error is a false positive, while type II error is false negative
Briefly stated, Type I error means claiming something has happened when it hasn't, While Type II error means that you claim nothing is happening when in fact something is.

What's a Fourier transform (FT)?
A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions.
A FT converts a signal from time to frequency domain.
The FT estimate the amplitudes, and phases to the given signal.

What is Neuron: An artificial neuron is a mathematical function. It takes one or more inputs that are multiplied by weights and add together. This value is then passed to a non-linear function, known as an activation function, to become the neuron's output.
It provides an output by applying the function on the inputs provided.

Neural networks, a beautiful biologically-inspired programming language/paradigm which enables a computer to learn from observational data.
NN composed of neurons, contains activation functions that makes it possible to predict non-linear outputs.

Deep learning, a powerful set of techniques for learning in neural networks. Deep learning is a subset of ML, and the code structures are arranged in layers. Each layer's input is the previous layer's output, which yields progressively higher-level features and defines a hierarchy.
A Deep Neural Network is just NN that has more than 1 hidden layer.
Deep learning frame works are MXNet, TensorFlow, Caffe, and PyTorch
Models: Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Multilayer Perception Neural Network (MLPNN), Long Short-Term Memory (LSTM), Generative Adversarial Network (GAN), Restricted Boltzmann Machine (RBM), and Deep Belief Network (DBN).
Applied to medical image processing, bioinformatics, social network filtering, fraud detection, image and speech detection, audio recognition, computer vision, customer relationship management, and many more fields.
It capable of training neural networks to deliver impressive results.

What is deep learning, and how does it contrast with other machine learning algorithms?
Deep Learning is a subset of machine learning that is concerned neural networks: how to use back propagation and certain principles from neuro science to more accurately model large sets of unlabeled or semi-structured data.
Deep learning represents an un supervised learning algorithm.

Tensorflow:
Tensorflow's name is directly derived from its core frame work tensor.
A tensor is a vector or matrix of n-dimensions that represents all types of data.
A tensor hold identical data type with a shape. The shape of the data is the dimensionality of the matrix or array.
Tensorflow is a opensource library.

TensorBoard: is the interface used to visualize the graph and other tools to understand, debug, and optimize the model.

Scikit-Learn: is an opensource library for machine learning built on top of Scientific Python (SciPy) in Python. It provides a several tools for ML and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistence interface in Python, and this library written in Python.

PyTorch: is a open source ML library based on Torch library, used for applications such as computer vision, and NLP.

Keras: Keras is a open source software library written in Python and capable for running on top of either TensorFlow or Theano. Using in computer vision and deep learning models.

Artificial Neural Network (ANN): ANN, is a group of multiple perceptions/neurons at each layer. (or) multilayer perceptional architecture.
ANN is also known as Feed-Forward Neural network because inputs are processed only in the forward direction. ANN consists of 3 layers, Input, hidden and output layers. The input layer accepts the inputs, the hidden layer processes the inputs, and the output layer produces the result. Essentially, each layer tries to learn certain weights.

A single perception (or neuron or node) can be imagined as a Logistic Regression.
ANN is being deployed across industries such as healthcare, financial markets, meteorology, and learning management systems among others.

Convolution Neural Network ( ): Multilayer perceptional architecture and it heavily used in the field of computer vision. It consists of convolution layers, activation layers, pooling layers, fully connected layers, and output or normalization layers.
Computer vision: medical image analysis, image recognition and face detection.

Recurrent Neural Network (RNN):
RNN is a multi-layer neural network, processing sequential data and identifying patterns in data, such as text, speech or videos.
RNN heavily used in Natural Language Processing (NLP) and speech recognition.
RNNs are designed to recognize patterns in sequences of data, such as text, genomes, handwriting, the spoken word, numerical time series data.
Language modelling and translation, Machine translation, speech recognition
RNN-LSTM

Natural Language processing (NLP): sentence modeling and search query retrieval. It is a branch of Artificial Intelligence (AI). NLP understand and translate the human language.
NLP tools and libraries: NLTK, Spacy, Apache, Genism, textblob library.
Recently started the NLP information retrieval (IR) and best tool kit is **wordnet**
(see writeup more details)
NLP metrics: BLEU score, perplexity
Perplexity: is the inverse probability of the test. $PP(w) = 2^{-l}$
Where $l = 1/N \log P(w_1, w_2, \ldots w_n)$

Face recognition system is available in Keras ie. FaceNet is available in Keras.

NLP pipeline: (images and scanned documents)
Categorizing scanned documents
Sentence segmentation: Document containing lots of text into sentence
Converting lowercase: Don't want to differentiate words based on their case
Removing the URL address: removal https, punctuations spl. Characters, and emojis
Removal stop words: commonly occurring words (in, a, the, and, an an, etc.)
Stemming: stemmer algorithm work by cutting off the end or beginning of the word.
(or) reducing the inflection words (playing, played, plays- root word is play)

Lemmatization: It will return the dictionary form of a word that can find in a dictionary.
Both stemming/lemmatization are word normalization techniques.

Tokenization: splitting paragraph, sentence, phrases into smaller units or words, and each smaller unites are tokens.
Bag of Words: model represents, turn arbitrary text into fixed length of vectors by counting how many times each word appears. This process is often referred to as vectorization.
TF-IDF: Term Frequency – Inverse Document Frequency
Is a numerical statistic and it reflect how important a word in a document or corpus.
TF: It is a measure of how frequently a term (t) appears in a document (d).
$Tf_{t,d} = n_{t,d}$ /number of terms in the document
Inverse Document Frequency: is a measure of how important of a term is
Idf_t = log(number of documents/number of documents with term 't')

Topic modelling: Latent Dirichlet Allocation (LDA) – probabilistic model (see page 22)

Back propagation: The weights of a neural network are updated through this backpropagation algorithm by finding the gradients.

Back propagation is an algorithm commonly used to train neural network.
The method calculates the gradient of the error function w.r.t the neural network's weights.
The 'backwards' part of the name stems from the fact calculation of the gradient procedure backwards through the network, with the gradient of the final layer weights being calculated first and the gradient of the first layer of weights being calculated last.

What cross-validation technique would you use on a time series dataset?
Generally, I used standard k-folds cross-validation
K-fold cross-validation (CV) would use on a time series data set.
K-fold cv splits into 'K' number of sections/groups. Where each group a s a testing set at some point.
Ex: If k=5 then it is 5-old cv. Here, the dataset split into 5 groups. In the 1st iteration, the first group is used to test the model and rest are used to train the model.

What is Imbalanced data?
In classification problem where the classes are not represented equally.
Ex. You may have a 2-class (binary) classification problem with 100 instances (rows). A total 80 observations are labeled with class-1 and 20 observations are labeled with class-20. It means 80:20 corresponds to 4:1.

How would you handle an imbalanced dataset?
Collect more data to even the imbalances in the dataset
Resample the dataset to correct for imbalances.
Use the right evaluation metric
Try different algorithms

When should you use classification over regression?
Classification produce discrete values (0 and 1) and dataset to strict categories, while regression gives you continuous results.
Classification involves the identification of values that lie in specific group.

Name an example where ensemble techniques might be useful?
Ensemble algorithms are those which combines more than one algorithm of same or different kind for classifying objects.
For example, running prediction over Naïve Bayes, SVM, and Decision Tree and then taking vote for final considering of class for test object.
They typically reduce overfitting in models and make the model more robust.
(or)
Training multiple models with different parameters to solve the same problem.
Ex. Of ensemble methods: bagging, boosting, the bucket of model method.

Bagging and Boosting
Both are ensemble methods to get N learners from 1 learner.
Both generate several training data sets by random sampling.
Both are good at reducing variance and provide higher stability.

Bagging	**Boosting**
Aim to decrease variance, not bias	Aim to decrease bias, not variance
Each model receives equal weight	Models are weighted according their performance
Each model built independently	New models are influenced by performance of previous built models
Random Forest	Gradient boosting
Random	Sequential

Bagging decreases variance, boosting decreases bias.
Under-fitting: means the model has low variance, high bias
Over-fitting: means the model has high variance and low bias
Boosting is more vulnerable to overfitting.

Which is more important to you: model accuracy or model performance?
Model accuracy is only a subset of model performance.

How do you ensure you're not overfitting with a model?
Using simpler model: reduce the variance by taking into account fewer variables and parameters.
Use cross-validation techniques such as k-fold cross-validation.
By reducing variance and adding some bias to the model.
Use regularization techniques such as LASSO or Ridge that penalize certain model parameters if they're likely to cause overfitting.

Cross-Entropy
Quantifies the difference between two probability distributions.

What evaluation approaches would you work to gauge the effectiveness of ML model?
Measure such as: the accuracy, F1-score, and the confusion matrix.

How would you evaluate a logistic regression model?
Classification accuracy, Area under ROC curve, confusion matrix, and classification report.

Different evaluation metrics are:
Classification accuracy, logarithmic accuracy, confusion matrix, Area under curve (AUC), F1-score, mean absolute error, mean squared error.

Classification accuracy:
It is the ratio of number of correct predictions to the total number of input samples.

Loss function and cost function:
A loss function is for a single training example. It is also sometimes called an error function.
(or) The loss functions are used to determine the error between the output of our algorithms and the given target value. In layman's terms, the loss function expresses how far off the mark our computed output is.
Loss functions are used in optimization problems to minimize the loss.

A cost function, is the average loss over the entire training dataset. The optimization strategies aim at minimizing the cost function.

Logarithmic or log loss:
log loss measures the performance of a classification model, and loss range [0,infinity). Log loss near to 0 indicates higher accuracy, whereas if the log loss is away from 0 then it is indicates lower accuracy.

What's the Kernel trick and how is it useful?
Using kernel trick enables us effectively run algorithms in a higher dimensional space with lower-dimensional data.

How would you build a data pipeline?
Select the problem, collect the data, preprocess, build the model, and fit the model. Host the model and pipelines into Google Cloud or AWS or Azure.

Multivariate Regression analysis:
Multivariate regression is a technique that estimates a single regression model with more than one outcome variable.

Multivariate analysis of variance (MANOVA) is an extensive of the univariate analysis of variance (ANOVA). In an ANOVA, we examine for statistical differences on one continuous dependent variable by an independent grouping variable. The MANOVA extends this analysis by taking in to account multiple continuous dependent variables.

Multi collinearity: more than two variables in a multiple regression model are highly linearly related or highly correlated.
Multi collinearity detect by the correlation matrix
Multi collinearity is a problem because two regressors or variables are perfectly correlated, their coefficients will be difficult to calculate and interpret.
Remove highly correlated predictors/variables from the model
Use Partial Least Squares Regression (PLS) or Principal Component Analysis (PCA) to cut the number of predictors to a smaller set of uncorrelated components.

Principal Component Analysis (PCA):
PCA is an unsupervised algorithm used for dimensionality reduction in ML.
(high dimensionality = datasets have large number of features)
Create correlation matrix or covariance matrix for all the desired dimensions
Calculate Eigenvectors that are the principal component (PC1, PC2…) and respective eigenvalues that indicates the magnitude of variance.
Arrange eigenpairs in decreasing order of respective eigenvalues and pick the value which has the maximum value, this is the first principal component (PC1).

Eigen vectors:
Eigen vectors are unit vectors, which means that their length of the magnitude is equal to 1.0
Eigen value equation A.V = lambda . V
A is the parent square matrix that we are decomposing, V is the **eigenvector** of the matrix, and lambda represents the **eigenvalue scalar**.
Eigenvectors and eigenvalues are used to reduce noise in data. They also eliminate features that have a strong correlation between them and also help in reducing over fitting. If negative value indicates the reverse direction.

Support Vector Machine (SVM):
SVM is supervised machine learning algorithm.
Which can be used for classifier tasks.
Support vectors are simply the co-ordinates of individual observation.
SVM is classifier, which best segregates/separate the two classes (hyper-plane/line).
Support vectors are the data points that lie closest to the decision surface (or hyperplane). It calculates the z^2=x^2 + y^2 (equation of a circle) and that is distance between the data point to the center. (or)
It uses the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

Long short-term memory (LSTM):
LSTM is an artificial Recurrent Neural Network (RNN) architecture in the field of deep learning.
LSTM networks are well-suited to classifying, processing, and making predictions based on time-series data.
It can be beneficial to allow the LSTM model to learn the input sequence both forward and backwards and concatenate both interpretations.
LSTM unit is composed of a cell, an input gate, an output gate and forget gate.
Important tasks that include language modeling, speech recognition, and machine translation.
The sequence-to-sequence LSTM, also called encoder-decoder LSTMs. The Encoder-Decoder LSTM was developed for Natural language processing (NLP) problems where it demonstrated state-of-art performance, specifically in the area of text translation called statistical machine translation.

Encoder-Decoder LSTM or GRU cells architecture: The architecture is composed of two models. One for reading the input sequence and encoding it into a fixed-length vector, and a second for decoding the fixed-length vector and outputting the predicted sequence. Encoder-Decoder LSTM designed specifically for seq2seq problems.

Applications of Encoder-Decoder LSTMs: Machine translation (ex. English to French translation of phrases).
Image Captioning (ex. Generating a text description for images)
Conversational modeling (ex. Generating answers to textual questions)

What is the difference between LSTM and GRU (Gated Recurrent Unit): GRU's has two gates one is update gate and reset gate, while LSTM has three gates are input gate, output gate, and forget gate.
GRU is improved version of standard recurrent neural network. Basically, update gate and reset gates are two vectors which decide what information should be passed to the output.

Name Entity Recognition (NER): It is the task of identifying and categorizing key information (entities) in text.
An entity can be any word or series of words that consistently refer to the same thing.
Two steps are detect a named entity, and Categorize the entity.

Reinforcement Learning (RL):
RL is the training of ML models to make sequence of decisions.
RL is a part of deep learning methods, it is taking suitable action to maximize reward in a particular situation.

Recommender Systems: is collaborative filtering. The amazon was the first company to leverage item-to-item collaborative filtering.
Content-based methods: uses attributes of items/users, recommended items similar to those liked by the user in the past.

Collaborative filtering methods: recommended items liked by similar users, enable exploration of diverse content.
You can find large scale recommender systems in retail, video on demand, or music streaming.

Survival analysis: is a statistical method that aims to predict the time to an event, such as death, the diagnosis of a disease or failure of a mechanical part.
To compare time-to-event between two or more groups.
To assess the relationship of co-variables to time-to-event.

Churn model: is a mathematical representation of churn impact on your business. Customer churn prediction can help you see which customers are about to leave your service so you can develop proper strategy to re-engage them before it is too late.

Why use XGBoost?
XGBoost stands for eXtreme Gradient Boosting
XGBoost is hierarchical decision-tree based machine learning algorithm.
It can use both classifier or regression tasks.
Boosting is a sequential technique works on the principal of ensemble method.
XGBoost is an ensemble method (ensemble methods to get N learners from 1 learner)
Execution speed.
Model performance

Model parameters:
Model parameter is a variable of the selected model which can be estimated by fitting the given data to the model.
Ex: x is independent variable, and y is the dependent variable. The objective is to fit a regression line to the data. This line (the model) is then used to predict the y-value for values of x.

Hyper parameters
A hyperparameter is a parameter that is set before learning process begins.
These parameters are tunable and can directly affect how well a model trains.

Learning rate, number of epochs, hidden layers, activation functions, number of branches in decision tree, number of clusters in a clustering algorithm (k-means clustering)

Linear programming: is suitable for solving linear optimization problems. It is a mathematical method (and associated algorithms) for maximizing or minimizing the function.

Dynamic programming: suitable for non-linear optimization problems.
Solving the problems by breaking them down into simpler problems. It is essentially a smart recursion.

Logistic regression:
Logistic regression is a supervised ML algorithm
Generally, it is used in binary classification algorithm
Ex: predicitions for yes/no, true/false A/B

Linear regression:
Linear regression is a supervised ML algorithm.
It is used in regression tasks.
Variables must have linear relationship, linear independence of variables.
Residuals must be normally distributed,
Data must be homoeskapacity, and no multi collinearity.

Simple linear regression:
SLR defines the relationship between single dependent variable and single independent variable.

Regression analysis
Deals with the one dependent variable and one or more independent variables.

Supervised and un supervised learning algorithms:
The majority of Machine learning algorithms are supervised algorithms.
Y = f(x)
You have a new input data (x) and predict the output variable (y).
Supervised algorithms are labelled.
Supervised algorithms are further grouped into classification and regression groups.

Unsupervised algorithms: all data is unlabeled data
You have only input data (x) and no corresponding output variable.
Clustering (k-means clustering, DBSCAN, and mean-shift clustering).

Statistical bias:
Statistical bias is defined as the difference between the parameter to be estimated and the mathematical expectation of the estimator.

Pickle:
We will save our trained model to the disk using pickle library.
Pickle is used to serializing and de-serializing structure (if the lengthy program time)

API: Application Programming Interface
Flask: Flask is an API of Python that allows us to build up web applications.
(or) Flask is a web framework, it's a Python module that lets you develop web applications easily.

Demographic data:
Demographics is the study of a population based on factors such as age, race, and sex.

Data Analysis:

1. Qualitative analysis is done through observations
2. Quantitative analysis done through surveys and experiments.

Data analytics process: Business problem, Data, preprocessing, model selection, train model, test model, metrics, results, and deploying the model.

Data profiling:
Data profiling is a process of examining the data from an existing source and summarizing information about the data. The flow is the ETL (extract, Transform, and Load Data).
Data profiling utilizes methods of descriptive statistics such as min, max, mean, mode, percentile, standard deviation, frequency, variation, and etc..

Data migration:
Data migration is the process of moving from one location to another, one format to another, or one application to another.
Generally, this is the result of introducing a new system or location for the data.

Data Integration:
Data integration involves combining data residing in different sources into a single, unified view.
Integration begins with the ingestion process, and includes steps such as cleaning, ETL mapping, and transformation.

What is database: A database is a collection of logically related information in an organized way so that it can be easily accessed, managed and updated. In addition, on database such as adding, updating and deleting data.

Graph database: It is a database, graph structures represents the data with nodes, edges, vertices and properties represented by key/value pairs.
Nodes can be labelled to be grouped. The edges representing the relationships have two qualities.

Relational database is a digital database, it is arranged in a column and rows with different (integer or floating) data points. The tables of columns and rows with a unique key of identifying each row. Rows are called records /observations, and columns are called variables.

Structured data: Data that is the easiest to search and organize, because it is usually arranged in rows (observations), and columns (features/variables) into pre-defined fields. Structure data is managed using Structured Query Language (SQL), and it is relational database.

Unstructured data: is non-relational database ie., not in rows and columns. Unstructured datasets are stored in data lakes, NoSQL databases. The examples are text, audio, and videos.

Difference between PySpark and Spark:

PySpark:

A tool to support Python with spark
Supported by library called Py4j, which is written in Python
Developed to support Python in Spark
Understanding of Big data and Spark

Spark:

A data computational framework that handles Big data
Written in Scala. Apache Core is the main component
Works with other languages such as Java, Python, and R
Pre-requisites are programming knowledge in Scala and database.

Apache Hadoop or Hadoop:

The Apache Hadoop software library is a framework that handles Big data sets across clusters of computers using simple programming models.
It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
(or)
Hadoop is an open-source software framework that is used for efficiently store and process large datasets ranging from gigabytes (GB:10^9) to petabytes (PB: 10^{15} or one million GB) of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

Amazon SageMaker

An Amazon SageMaker notebook instance is a machine learning (ML) compute instance running the Jupyter Notebook App. SageMaker manages creating the instance and related resources. Use Jupyter notebooks in your notebook instance to prepare and process data, write code to train models, deploy models to SageMaker hosting, and test or validate your models.

Swagger UI: provides a display framework that reads an OpenAPI document and generates an interactive documentation website.

The **Simple Regression** model, relates one independent variable (predictor) and one dependent variable (response).
The **Multiple Regression** model, relates more than one independent variable (predictor) and one dependent variable (response).

The **Multivariate Regression** model, relates more than one independent variable (predictor) and more than one dependent variable (response), are linearly related. The multivariate technique

allows finding a relationship between variables or features. It helps to find a correlation between independent and dependent variables.

Learning rate: is tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of cost function.

How to configure the learning rate when training deep learning neural network
The weights of the neural network cannot be calculated using an analytical method. The weights must be discovered via an empirical optimization procedure called stochastic gradient descent (SGD).

One of the key hyperparameter to set in order to train a neural network is the learning rate for gradient descent. This parameter scales, the magnitude of our weight updates in order to minimize the network's loss function.
If your learning rate is too low, training will progress very slowly.

Optimizers: Adam, SGD, RMSProp are the best optimizers

Using **stochastic gradient descent** (SGD) optimizer algorithm with different learning rate (lr) schedules to compare the performances. SGD optimizer is available in Keras. The methods are: Time-Based decay, step decay, exponential decay
Time based decay (lr = lr0/(1+kt), lr and K are hyperparameter, and t is the iteration number).

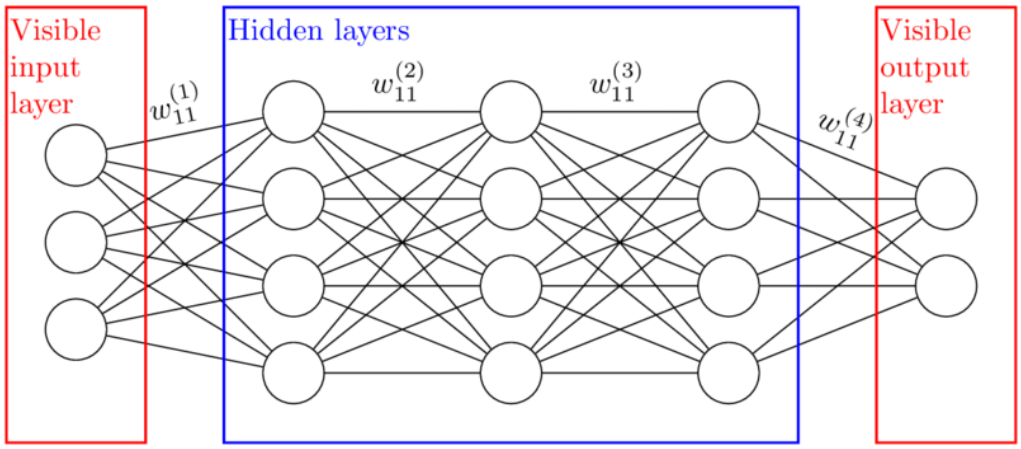
Batch size: The batch size defines the number of samples that will be propagate the through neural network.
Ex: Let's say you have 1050 training samples and you want to set up a batch size equal to 100. The algorithm takes the first 100 samples (1-100) and trains the network. Next, it takes the second hundred (101-200) and trains the network again. We can keep doing procedure until finish all the samples. In this case problem it may happen because 1050 not divisible by 100, and best way is batch size is 50 for this case.

Dataset of 2000 training samples into batch size is 200, so number batches (number of iterations) are 10 to complete 1 epoch.

One epoch is when an entire dataset is passed forward and backward through the neural network only once. Since one epoch is too big to feed to the computer at once we divide into smaller batch sizes.

Why we use more than one epoch:
Passing entire dataset in neural network is not enough. We need to pass the full dataset multiple times to the same neural network. We are using optimization techniques ie. Gradient descent which is an iterative process. So, updating the weights with single pass or one epoch is not sufficient.

Hidden layer: is located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through the activation function as the output.
Hidden layers perform nonlinear transformations of the inputs entered into the network.
One hidden layer is sufficient for the large majority of problems.
The larger the number of hidden layers in a neural network, the longer it will take for the neural network to produce the output and it can solve complex neural network problems.



How many neurons (nodes) in Input layer?
The number of neurons in an input layer is dependent on the shape of your train data.
Number of neurons (nodes) = number of training data features +1
One additional node is to capture the bias term.

How many neurons (nodes) in output layers
If your neural network is a regressor or classifier, the output layer has a single node (if you use sigmoid activation just a single node). If you use a probabilistic activation function such as softmax then the output layer has one node per class label in your model.

How do you select the hidden layers in neural network?
The number of hidden neurons (nodes) should be less than twice the size of the input layer.
The number of hidden neurons should be less than 2/3 the size of the input layer, plus the size of output layer

Chi-Square test: is commonly used for testing relationships between categorical variables.
$\chi^2 = \sum(X_i - E_i)^2 / E_i$
X=observed value

E = expected value

A/B testing: is a basic randomized control experiment. Compare the two variations of a variable to find which perform better.

Feature engineering techniques: is the process of transforming the raw data into features that better represent the underlying problem to the predictive models.
Imputations (missing data), handling outliers, binning, log transform, feature split, normalizations (scaling).

Advanced NLP techniques:
Encoder-decoder,
Embeddings (BERT, ELMO, ULMFit, Skip-Gram, Glove, and Word2Vec)
Sentiment analysis
Latent Dirichlet Allocation (LDA)
Latent Segment Analysis (LSA)

Embeddings: is just numerical representation of words.
RGB representation for colors.

Bag of Words (BOW): model is a representation that turns arbitrary text into fixed length of vectors by counting how many times each word appears.
This process is often referred to as Vectorization creates text to vectors

CBOW (Continuous Bag of Words): The way CBOW work is that it tends to predict the probability of a word given a context. A context may be single word or a group of words.

Skip-gram: is one of the unsupervised learning techniques to find the most related words for a given word.

Word2Vec model generates the embeddings that context-independent, ie., there is one vector (one numeric value) representation for each word. Different senses (meaning) of the word (if any) are combined into one single vector.

BERT: (Bidirectional encoder representations Transformer) this model generates embeddings that allows us to have multiple (more than one) vector (numeric) representations for the same word, based on context in which the word is used.
BERT embeddings are context-dependent
BERT is bi-directional LSTM. The architecture is encoder-decoder.
It looks at text left to right and right to left.
The key is multi headed attention, BERT3 is better than BERT
Ex: We went to river bank
I need to go to bank to make deposit.

Glove: is also context-independent like word2vec.

ELMO model generates embeddings. ELMO has both right to left and left to right simultaneously.

Sentiment Analysis:
It is a supervised ML algorithm
Sentiment analysis is the process of detecting positive or negative sentiment in text. (or)
Classifying opinions found in text into categories like “positive” or “negative” or “neutral”.

Sentiment analysis models focus on polarity (positive, negative, neutral) but also feelings and emotions (angry, happy, sad, etc..), urgency (urgent, not urgent) and even intentions (interested vs not interested).

Ex: “The story of the movie was very interesting and good narration”

Opinion owner: Audience
Object: movie
Feature: Story
Opinion: very interesting and good narration
Polarity: positive

Latent Dirichlet Allocation (LDA): LDA is a probabilistic model.
It is unsupervised algorithm
LDA is one of the most popular topic modeling methods.
LDA assumes documents are produced from a mixture of topics.
Those topics then generate words based on their probability distribution.
Every document is a mixture of topics.
Every topic is a mixture of words.

The topic modeling is a branch of unsupervised NLP. It is a probabilistic model, which is used a text document with the help of several topics. Doing in terms of clustering, instead of numerical features, we have collection of words that we want to group together in such way each group represents a topic in a document.
Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic.
It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

Image preprocessing steps are: Uniform aspect ratio: ensure the images have the same size and aspect ratio. Most of the neural network models assume a square shape input image, which means that each image needs to be checked if it is square or not. (or)
A fixed size must be selected for input images, and all images must be resized to that shape.
Aspect ratio of image is the proportional relationship of the width to the height. (ex. A image of 4x2 inch then the aspect ratio will be 2:1)
Image scaling/ image normalization: the image intensity range between 0 to 255 and these values are rescaling to 0-1.

Dimensionality reduction: Try to change the RGB (color) channels into a single gray-scale channel.
The augmentations are required, involve random rescaling, horizontal flips, perturbations to brightness, and contrast, as well as random cropping.
Bilateral filtering performs noise suppressing and smooths images while preserving the edges by replacing the intensity value at each pixel with weighted average of intensity values from nearby pixel.
Total variation denoising methods commonly enhance edges and outlines but suppress the details and perturbations.

The three types of pixel scaling/normalization
Pixel Normalization: scale pixel values to the range 0-1
(pixel values in the range between 0 and 255, and these values rescaling to 0-1)
Pixel centering: scale pixel values to have a zero mean
Pixel standardization: scale pixel values to have a zero mean and unit variance.
Simplified whitening, Local contrast normalization, and Local response normalization,

Optical Character Recognition (OCR): Machine printed or hand written ie. scanned documents, or pdf docs, or images are converted into machine readable text docs using OCR.
OCR as a process generally consists of several sub-process are:
Preprocessing of the Image, text localization, character segmentation, and post processing.
Tesseract is an open source text recognition (OCR) engine, and OpenCV.

YOLO vs SSD:
YOLO (You Only Look ONCE): an open source method of object detection that can recognize objects in images and videos.

SSD (Single Shot Detector) runs a convolutional network on input image only one time and computes a feature map.

NLP Machine Translation:
Machine Translation (MT) is a subfield of computational linguistics that is focused on translating text from one language to another.
Neural machine Translation (NMT) is the most powerful algorithm to perform this task. This based on encoder-decoder LSTM structure.
Encoder-decoder LSTM designed for specifically for seq2seq problems.

Language detection and translate one language to another:
For detect language for particular data. For this you can use a “langdetect” python package.
It support 55 languages.
For translate one language to another language use “google_trans_new” python package.
(see lang_detect.ipynb)

NLP Terminology:

Phonology: is the first level of Natural Language Understanding (NLU).
It deals with speech recognition and generation. Markov model is the best one. (or)
Is a branch of linguistics, which studies the manners of organization and usage of the speech sounds in natural language.
Majority of languages have about 30 phonemes, but some that have as few as 11 or as many as almost 150. The English language has about 43 phonemes.

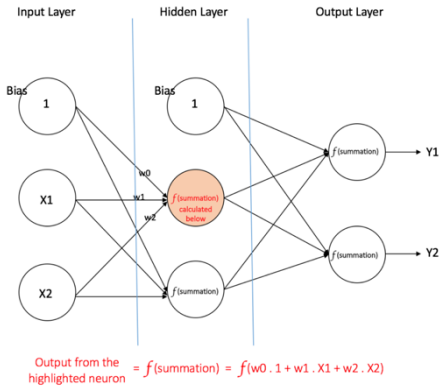
Morphology: It is a study of construction of words from primitive meaningful units.
Morphology is the study of the structure of words, specifically looking at roots, affixes, and parts of speech. A morpheme is the smallest unit of meaning within a word.
Morpheme:
It is a primitive unit of meaning in a language. (or)
A morpheme is the smallest unit of meaning within a word.
There are two types, lexical morphemes and grammatical morphemes.
Syntax:
It refers to arranging words to make a sentence. It also involves determining the structure role of words in the sentence and in phrases.

Semantics:
It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
Pragmatics:
It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.

Linguistics analysis:
Lexical analysis: It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language.
Lexical analysis is dividing the whole chunk of txt into paragraphs, sentences, and words.
Syntactic analysis (Parsing):
It involves analysis in the sentence for grammar and arranging words in a manner that shows the relationship among the words.
Semantic analysis:
It draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain.
Discourse Integration:
The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.
Pragmatic analysis:
During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge.
(https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm)

Number of neurons (nodes) = number of training data features +1 (bias)
The total neurons (nodes) in the above neural network
Input layers:2 nodes (X1 and X2) + (add 1 for bias)
Hidden layers: 2 nodes + 1 bias
Output layer: 2 (Y1, and Y2)

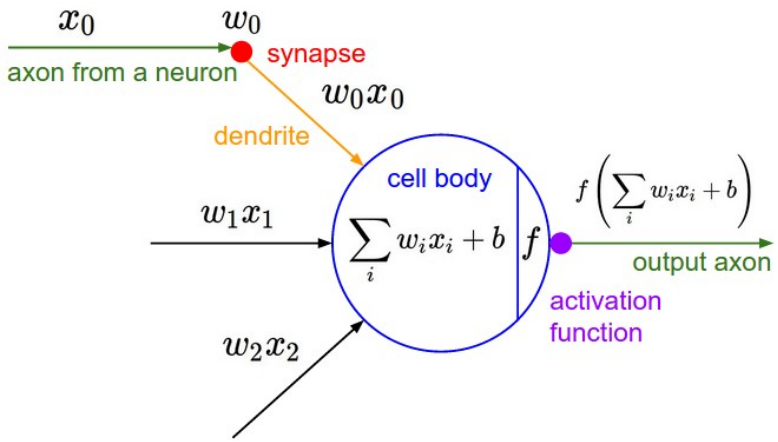
[(2*2)+(2*2)] = 8 weights and (1*2 +1* 2) =4 biases, the total is 12 learnable parameters
(or)
Simple: [(input nodes+1)*hidden nodes] + [(hidden nodes+1)*output nodes]
: [(2+1)*2] + [(2+1)*2] = 6+6 =12



Example:
Input: 3 nodes
Hidden: 2 nodes
Output: 1 node
Simple: [(input nodes+1)*hidden nodes] + [(hidden nodes+1)*output nodes]
: [(3+1)*2] + [(2+1)*1] = 8 + 3 =11
Total learnable parameters: [(3+1) *2] + [(2+1) *1] = 8+3 =11
(or)

Simple: [(3*2)+(2*1)]+[(1*2)+(1*1)] = 8+3=11

Reverse engineering (or sometimes back-engineering) is a process that is designed to extract enough data from a product and then to be able to produce that product. It may involve moving to creating a product from scratch or from pre-developed components. Three main stages: Implementation recovery, Design recovery, and analysis recovery.



An example code for forward-propagating a single neuron might look as follows:

```
class Neuron(object):
  # ...
  def forward(self, inputs):
    """ assume inputs and weights are 1-D numpy arrays and bias is a number """
    cell_body_sum = np.sum(inputs * self.weights) + self.bias
    firing_rate = 1.0 / (1.0 + math.exp(-cell_body_sum)) # sigmoid activation functior
    return firing_rate
```

Analytical methods:
Data mining is the process of identifying sequences, relationships, or anomalies in large amounts of datasets.
Cluster analysis, Regression analysis, time series analysis, and sentiment analysis.

Predictive analytics: makes predictions about certain unknown in the future.

In the sense of ML, from data mining, predictitve modelling that analyze current and historical facts to make predicitions about future or otherwise unknow events.

Predictive modeling is the process, test and validate a model. The models are Logistic regression, Random Forest, and Decision trees.

Cash forecasting: Supervised ML algorithm.
Forecasting daily cash flows fir the next 3 months using daily incoming and outgoing cash flows.

Price optimization: is the process of finding that pricing or maximizing price against the customers willingness to pay.
Price optimaztion models are mathematical programs that calculate how demand varies at different price levels. The resulting data is then combined with information on costs and stock levels to recommend prices that will improve the profits.
Coolect historical data, include product volumes, the company's prices and promotions.

Difference between PNG and JPEG?
JPEG uses lossy compression algorithm, whereas PNG uses lossless compression algorithm and no image data loss is present in PNG format.

Model Validation: is the process of evaluating a trained model on test data set. This provides the generalized ability of trained model.
Split the data into training and test datasets. Define metrics for which model is getting optimized.

Model Serving:
TensorFlow serving is a flexible, high-performance serving system for ML models, designed for production environments.
TensorFlow serving makes it easy to deploy new algorithms and experiments, while keeping the same server architecture.
(or)
Databricks ML flow Model Serving provides a turnkey solution to host ML models as REST endpoints that are updated automatically.

Model monitoring:
The final phase, where we ensure our model is doing what we expect it to in production.

Model Retraining:
Retraining typical needed when a model quality degrades below a predetermined accuracy value.
If you see the accuracy of your model degrading over time, use the new data, or a combination of the new data and old training data to build and deploy a new model.

Deployment: Getting the model into production where it can start adding value by serving predictions. Typical artifacts are application programming interface (APIs) for accessing the model.

Kubeflow: as ML toolkit for kubernetes. Kubeflow is dedicated to making ML on kubernetes easy, portable, scalable.

Zipline: is a pythonic algorithmic trading library. It is an event-driven system that supports both back testing and live trading.
Currently used as the back testing and libe-trading engine powering Quantopian

Pyfolio: is an open source python library for analysis of the performance and risk of financial portfolios.

Pysf: Supervised forecasting is the ML task of making predictions for sequential data like time series.

PyFlux: is a library for time series analysis and prediction.

Pyramid: Python web frameworks.
The world python web frameworks are Django, Flask, Pyramid, Tornado, Bottle, Diesel, Pecan, Falcon, and many more.
Django and Pyramid both come with bootstrapping tools built in.
Bootstrapping is one of the most popular front-end frameworks, and it contains some amazing CSS classes for User Interface (UI) development.

Financial modeling: ML is a branch of AI that uses statistical models to make predictions. ML algorithms are used to detect fraud, automate trading activities, and provide financial advisory services to investors.
Ex: High-frequency Trading, Fraud detection, Loan/insurance underwriting, risk management, and chatbots.

Statistical techniques: Mean, variance, skewness, kurtosis, Linear regression analysis, hypothesis testing, Analysis of variance (ANOVA).

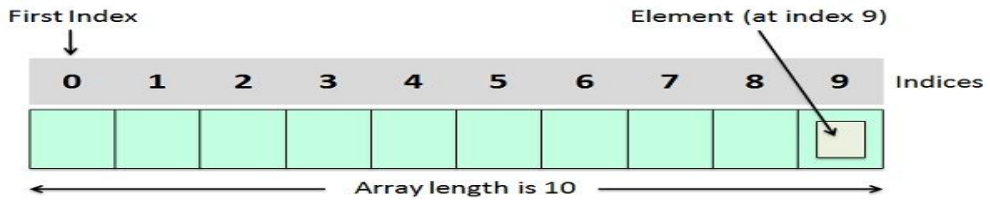
Statistical techniques for data scientist: Linear regression, classification, Resampling methods, subset selection, Regression techniques, Dimension reduction, nonlinear methods, Tree-based methods, Support vector machines, clustering, and unsupervised methods.

Scrum/Agile: Scrum is a framework for project management that emphasizes teamwork, accountability and iterative progress toward a well-defined goal.
Agile is a project management philosophy which utilizes a core set of values or principles.
Scrum is a specific agile methodology that is used to facilitate a project.

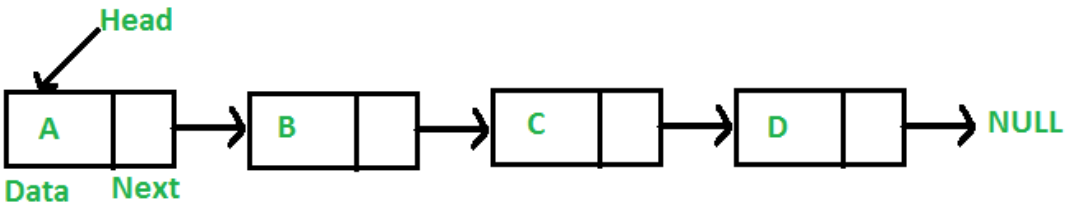
Big Data: is a collection of data that is huge in size and increasing exponentially with time.
Big Data analytics include stock exchanges, social media, jet engines, etc…

Financial engineering: is a multidisciplinary field involving financial theory, engineering methods, mathematical tools.
Financial engineering consists of converting financial theories into practical applications in the financial world.

What are some differences between a linked list and an array?
An array is a collection of elements of a similar data type.
Linked list is an ordered collection of elements of the same type in which each element is connected to the next using pointers.
Array elements can be accessed randomly using the array index.
Linked list not possible to access randomly.



Array



Linked List

Describe a hash table?
A hash table is a type of data structure that stores key-value pairs.
The key is sent to a hash function that performs arithmetic operations on it.

Which data visualization libraries do you use?
Popular tools:
Python's seaborn, Plotly, and matplotlib
R's ggplot
Tableau

How does XML and CSVs compare in terms of size?
XML is more readable than CSV
XML is natively supported by .net framework
XML stands for extensible markup language
CSV stands for comma separated values

What are the data types supported by JSON?
JSON stands for Javascript Object Notation
JSON datatypes: strings, objects, arrays, Booleans, and null values.

- **ResNet50 Architecture**
- The Architecture has 4 stages.
- Input image having height, width as multiples of 32 and 3 channel width.
- Input image size: 224x224x3
- Initial convolution and max-pooling using 7x7 and 3x3 kernel sizes.
- Afterward, Stage 1 of the network and it has 3 residual blocks containing 3 layers each.
- The size of the kernel in all 3 layers in Stage 1 are 64, 64 and 128.
- Residual block is performed with stride 2, hence the size of the input will be reduced but channel width will be doubled.
- Finally, the network has an average pooling layer followed by a fully connected layer having 1000 neurons (ImageNet class output).
- https://cv-tricks.com/keras/understand-implement-resnets/

Neural net works are multi layer perceptional architecture
Convolution layer, Activation layer, pooling layer, fully connected layer, and output layer.
VGG16 model Architecture:
Deep learning image processing (classification) – Keras, PyTorch

- 16 convolutional and max pooling layers (3x3)
- 3 dense layers for fully connecting layers, and
- An output layer of 1000 nodes.

Image preprocessing steps are: **Uniform aspect ratio**: ensure the images have the same size and aspect ratio. Most of the neural network models assume a square shape input image, which means that each image needs to be checked if it is square or not.
(or)
A fixed size must be selected for input images, and all images must be resized to that shape.
Aspect ratio of image is the proportional relationship of the width to the height. (ex. A image of 4x2 inch then the aspect ratio will be 2:1)
Image scaling/ image normalization: the image intensity range between 0 to 255 and these values are rescaling to 0-1.
Dimensionality reduction: Try to change the RGB (color) channels into a single gray-scale channel.
The augmentations are required, involve random rescaling, horizontal flips, perturbations to brightness, and contrast, as well as random cropping.
Bilateral filtering performs noise suppressing and smooths images while preserving the edges by replacing the intensity value at each pixel with weighted average of intensity values from nearby pixel.

Total variation denoising methods commonly enhance edges and outlines but suppress the details and perturbations.

The three types of pixel scaling/normalization
Pixel Normalization: scale pixel values to the range 0-1
(pixel values in the range between 0 and 255, and these values rescaling to 0-1)
Pixel centering: scale pixel values to have a zero mean
Pixel standardization: scale pixel values to have a zero mean and unit variance.
Simplified whitening, Local contrast normalization, and Local response normalization,

Optical Character Recognition (OCR): Machine printed or hand written ie. scanned documents, or pdf docs, or images are converted into machine readable text docs using OCR.
OCR as a process generally consists of several sub-process are:
Preprocessing of the Image, text localization, character segmentation, and post processing.
Tesseract is an open source text recognition (OCR) engine, and OpenCV.

Deep Neural Network Object detection 2D/3D: R-CNN and YOLO
R-CNN performs segmentation based on the results of object detection.

R-CNN extract many candidates and then calculates its characteristics through CNN.
Finally, it classifies each region using a specific linear classifier (SVM)
R-CNN built on traditional classification networks, AlexNet, VGG, GoogleNet, and ResNet.

YOLO: You Only Look Once: based CNN family of models for object detection and most recent version YOLOv3.
Open source library
YOLOv3 to perform object localization and detection on new image.
YOLOv3 multi layer predictions, a better backbone classifier, and more.

Apache TVM: is open source ML compiler framework for CPU and GPU, and ML accelerator.

OpenVINO stands for Open Visual Interface Neural Network Optimization.
Toolkit provided by Intel to facilitate faster interfence of deep learnig models.

TensorRT: High-performance neural network inference optimizer and runtime engine for production deployment (NVIDA TensorRT)
It supports both CPU and GPU inference.

Segmentation: the process of dividing customers into different groups based on their behavior or characteristics.

What is Amazon Web Service (AWS)
AWS (Amazon Web Services) is a platform to provide secure cloud service, database storage, offerings to compute power, content delivery, and other services to help business level and develop.
AWS, which is a cloud computing platform. It provides cloud services in the form of small building blocks, and these blocks help create and deploy various types of applications in the cloud.

What are the main components of AWS?
Simple Email service, Route 53 (It's a DNS web service), Simple Storage Service (S3), Elastic compute cloud (EC2), Elastic Block Storage (EBS), and Cloud watch. (DNS: Domain Name Systems)

List different ways to access AWS
Three different ways: console, SDK (software Development Kit), and CLI (Command Line Interface).

Which AWS region is the cheapest?
The US standard is the cheapest region (N. Virginia); it is also most established AWS region.

What are the most popular AWS services?
Amazon S3, Amazon Glacier, Amazon EC2, Amazon CloudFront, Amazon EBS, AWS Lambda, and Amazon Kinesis.

What are the EBS volumes?
EBS is Elastic Block Storage and it can attach to the instances. The EBS volumes will store the data even if you stop the instances.

What is deploy in AWS?
Code deploy is a deployment service that automates application deployment to Amazon EC2 instances, serverless Lambda functions or amazon ECS services.

Explain about AWS Lambda?
AWS Lambda is a computational service that enables you to run code without maintaining any servers. It automatically executes the code whenever needed. Lambda enables you to run the code virtually for any code of application without managing any servers. You are required to pay for the time that you have used it for.

Explain about S3:

S3 stands for Simple Storage Service. It is used for storing and retrieving data at anytime and anywhere on the web. S3 makes web-scale computing easier for developers.

AWS EMR (Elastic Map Reduce)?
EMR is a big data platform for data processing, interactive analysis, and machine learning using open source frame works such as Apache Spark, Apache Hadoop.

What is the maximum number of S3 buckets you can create?
100

How many total amazon Virtual Private Cloud's (VPC's) per account/region and subnets per VPC can you have?
VPCs per Region: 5
Subnets per VPC: 200

What is IAM?
IAM stands for Identity and Access Management. In this public IP address will change but Elastic IP address not changeable.

What is AMI?
It stands for Amazon Machine Image. The AMI contains essential information required to launch an instance, and it is a copy of AMI running in the cloud.

What is an EIP?
The Elastic IP address (EIP) is a static Ipv4 address offered by AWS to manage dynamic cloud computing services. Connect your AWS with EIP so that if you want a static IPv4 address for your instance, you can be associated with the EIP which enables communication with the internet.

Explain about DynamoDB:
If you want to have a faster and flexible NoSQL database, then the right thing available is DynamoDB, which is flexible and efficient database model available in AWS.

What is Glacier?
The Glacier is an online web storage service that provides you with low cost and effective storage with security features for archival and data backup. You can store the information effectively for months, years, and even decades.

Which one of the following is structured data store that supports indexing and data queries to both EC2 and S3?
Simple DB

What is an ELB?
Elastic Load Balance (ELB) is a load balancing service offered by AWS. It distributes incoming resources and controls the application traffic to meet traffic demands.

Mention the security best practice for Amazon EC2

Security and network, storage, resource management, and recovery and backup

What are the types of load balancers in EC2?
There are three types
Application Load Balancer: to make routing decisions at the *application* layer.
Network Load Balancer: to make routing decisions at the *transport* layer. It handles millions of requests per second.
Classic Load Balancer: to make routing decisions wither *application layer or transport layer*.

Explain what is T2 instance?
T2 instance is one of the low-cost Amazon instances that provides a baseline level of CPU performance.
What is Glue is a fully managed ETL (Extract, Transfer, and Load) service that makes it simple and cost effective to categorize your data, clean it, and move it data services and data streams.

What is Elastic Beanstalk?
Elastic Beanstalk is the best service offered by AWS for deploying and managing applications. It assists applications developed in Java, .Net, Node.js, PHP, Ruby, and Python. When you deploy the application, Elastic beanstalk built the selected supported platform versions and AWS services like S3, SNS, EC2, cloud watch, and auto scaling to run your application.
Jenkins is also for deploying managing application.

Mention a few benefits of the Elastic beanstalk?
Easy and Simple: Elastic Beanstalk enables you to manage and deploy the application easily and quickly.
Autoscaling: Beanstalk scales up or down automatically when your application traffic increase or decrease.
Developer productivity: Developers can easily deploy the application without any knowledge.
Cost-effective: No charge for Beanstalk. Charges are applied for the AWS service resources.
Customization:
Management and updates

What is meant by cloud watch?
Cloud watching is a monitoring tool in AWS with which you can monitor different resources of your organization. You can have a look at various things like health, applications, network etc..

How many types of cloud watches do we have?
We have two types of cloud watches: essential monitoring and detailed monitoring

What would be the minimum and maximum size of the individual objects that you can store in s3?
The minimum size of the object that you can store in S3 is 0 bytes, and the maximum size of the individual object that you can save is 5TB.

Explain the various storage classes available in S3
Glacier, standard frequency accessed, standard infrequency accessed

How many IP addresses are allowed for each account in AWS?
For each AWS account 5 VPC (Virtual Private Cloud) elastic addresses are allowed.

Give few examples of DB (database) engines that are used in AWS RDS
Oracle DB, MS-SQL DB, MYSQL DB, Postgre DB, and Maria DB.

Difference between block storage and file storage?
Block storage: it functions at a lower level and manages the data as set of blocks.
File Storage: The file storage operates at a higher level or operational level and manages data in the form of files and folders.

Difference between Elastic Block Storage (EBS) and S3 (Simple Storage System)

EBS	**S3**
Highly scalable	Less scalable
It is a block storage	it is an object storage
EBS is faster than S3	S3 is slower than EBS
It supports the file system storage	It supports web interface
User can access EBS only via the given EC2 instance	Anyone can access S3; it is public instance

How would you build a data pipeline?
Select the problem, collect the data, preprocess, build the model, and fit the model.
Host the model and pipelines into Google Cloud or AWS or Azure.

What is Amazon SageMaker?
Amazon SageMaker is a fully managed machine learning service. Data Scientists and developers can quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment.

Amazon Redshift: Redshift makes it simple and cost effective to run high performance queries on petabytes of semi-structured and structured data.
In redshift, each compute node is partitioned into slices, and each slice receives part of the memory and disk space. Slices work in parallel to perform the operations.

What is Redshift?
Redshift is a big data product used as a data warehouse in the cloud. It is the fast, reliable, and powerful product of a big data warehouse.

Athena: is a serverless interactive query service to analyze data in S3 using SQL.
Athena is used to with large scale data sets.
Amazon Athena enables users to analyze data in Amazon S3 using SQL.

Kubernetes: allows us to deploy and manage containerized applications at scale.
Kubernetes manages clusters of Amazon EC2 compute instances and run containers on those instances with processes for deployment, maintenance, and scaling.

Model Serving:
TensorFlow serving is a flexible, high-performance serving system for ML models, designed for production environments.
TensorFlow serving makes it easy to deploy new algorithms and experiments, while keeping the same server architecture.
(or)
Databricks ML flow Model Serving provides a turnkey solution to host ML models as REST endpoints that are updated automatically.

Model monitoring:
The final phase, where we ensure our model is doing what we expect it to in production.

Databricks: big data processing platform created by Apache Spark. Databricks was created for data scientists, engineers, and analysts to help users integrate fields of data science, engineering and the business behind them across the ML lifecycle.

AWS Databricks: is an industry-leading, cloud-based data analytics platform for data engineering, machine learning, and transforming massive quantities of data and exploring the data through ML models.

Model Retraining:
Retraining typical needed when a model quality degrades below a predetermined accuracy value.
If you see the accuracy of your model degrading over time, use the new data, or a combination of the new data and old training data to build and deploy a new model.

Deployment: Getting the model into production where it can start adding value by serving predictions. Typical artifacts are application programming interface (APIs) for accessing the model.

Docker helps us to create containers, and **Kubernetes** allows us to manage them at runtime.
Docker: is a open source platform for developing, shipping, and running applications. It makes easy for teams (developers and testing teams) to run applications in similar environment without any issues of dependencies or OS as it provides its own OS libraries.
Docker architecture: contains server, Rest API, command line interface (CLI).
Server: it can create and manage docker images
Rest API: instruct docker daemon what to do
CLI: is used to enter docker commands.
Docker container: is a executable image. You can create, start, stop, move or delete a container using Docker API or CLI.
Docker daemon (dockerd): listens for docker API requests and manages Docker objects such as images, containers, network, and volumes.

Docker Registry: stores Docker images. The same image might have multiple different versions identified by their tags.

Kubernetes: as a container orchestration tool.
It can handle container deployment, scaling, and load balancing of containers.

Kubeflow: as ML toolkit for kubernetes. Kubeflow is dedicated to making ML on kubernetes easy, portable, scalable.

A CI/CD pipeline:
A CI/CD pipeline automates the process and continuous monitoring throughout the lifecycle of a software product.
This pipeline is responsible for building codes, running tests, and deploying new software versions.

CI: continuous integration is a software development method. Integrates the code into a shared repository. It uses automated verifications for the early detection of problems.

CD: continuous delivery is the automated delivery of completed code to environments like testing and development.
Continuous deployment: is the next step of continuous delivery. Every change that passes the automated tests is automatically placed in production.

Jenkins: Jenkins is a open-source automation tool used to build and test software projects.
This tool more convenient for developers to integrate changes to the project.
Jenkins achieves continuous integration (CI) with the help of plugins.

Azure is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through MS managed data centers.
Databricks ML is a comprehensive tool for developing and deploying machine learning models with Azure Databricks.
It includes the most popular machine learning and deep learning libraries, as well as MLflow, a machine learning platform API for tracking and managing the end-to-end ML lifecycle.
Create Databricks workspace, cluster, and notebook. Run code in a Databricks notebook wither interactively. Train a ML model using Databricks, and deploy a Databricks-trained ML model as a prediction service.
Create feature tables and access them for model training and inference.

Google Cloud Platform (GCP)
For my private Bigdata Machine Learning project purpose I used the Google Cloud platform (GCP).
AI platform supports data preparation (ingest, clean, feature engineering) using my bigdata sets.
For labelling training datasets I used Data labelling service provided by the GCP.
I build my model using cloud AutoML and easy-to-use graphical interface.
I use AI Platform Notebooks (hosted by Jupyter Notebooks) for building custom ML models using Tensorflow and Keras.

Databricks: big data processing platform created by Apache Spark. Databricks was created for data scientists, engineers, and analysts to help users integrate fields of data science, engineering and the business behind them across the ML lifecycle.

Docker: is a open source platform for developing, shipping, and running applications. It makes easy for teams (developers and testing teams) to run applications in similar environment without any issues of dependencies or OS as it provides its own OS libraries.

Kubernetes: as a container orchestration tool.
It can handle container deployment, scaling, and load balancing of containers.

Kubeflow: as ML toolkit for kubernetes. Kubeflow is dedicated to making ML on kubernetes easy, portable, scalable.

Model monitoring:
The final phase, where we ensure our model is doing what we expect it to in production.

Model Retraining:
Retraining typical needed when a model quality degrades below a predetermined accuracy value.
If you see the accuracy of your model degrading over time, use the new data, or a combination of the new data and old training data to build and deploy a new model.

Deployment: Getting the model into production where it can start adding value by serving predictions. Typical artifacts are application programming interface (APIs) for accessing the model.

Scala: programming language. Scala supports both object oriented programming and functional programming. Scala runs of Java Virtual Machine (JVM).

AKKA: is free and open source toolkit. Akka is a toolkit and runtime for building highly concurrent, distributed, and fault tolerant event-driven applications on the JVM. AKKA can be used with both java and Scala.

Reactive Streams: is an initiative to provide a standard for asynchronous stream processing with non-blocking back pressure.

Apache Kafka: is a open-source distributed event streaming platform. It is an open source software platform developed by Apache, and written in Scala and Java.

AWS RDS (Relational Database Service): It is a web service running in the cloud. It easy to set up, and scale a relational database in the cloud.

AWS Elastic File System (EFS): EFS is a highly scalable file storage system designed to provide flexible storage for multiple EC2 (Elastic Cloud Compute) instances.

What is database: A database is a collection of logically related information in an organized way so that it can be easily accessed, managed and updated. In addition, on database such as adding, updating and deleting data.

Graph database: It is a database, graph structures represents the data with nodes, edges, vertices and properties represented by key/value pairs.
Nodes can be labelled to be grouped. The edges representing the relationships have two qualities.

Relational database is a digital database, it is arranged in a column and rows with different (integer or floating) data points. The tables of columns and rows with a unique key of identifying each row. Rows are called records /observations, and columns are called variables.

Structured data: Data that is the easiest to search and organize, because it is usually arranged in rows (observations), and columns (features/variables) into pre-defined fields. Structure data is managed using Structured Query Language (SQL), and it is relational database.

Unstructured data: is non-relational database ie., not in rows and columns. Unstructured datasets are stored in data lakes, NoSQL databases. The examples are text, audio, and videos.

SQL and NoSQL: SQL databases are relational, NoSQL databases are non-relational.
SQL databases use structured query language and have a predefined schema.
NoSQL databases have dynamic schemas for unstructured data.
SQL databases are vertically scalable
NoSQL databases are horizontally scalable.

The term **schema** refers to the organization of data as a blueprint of how the database is constructed.

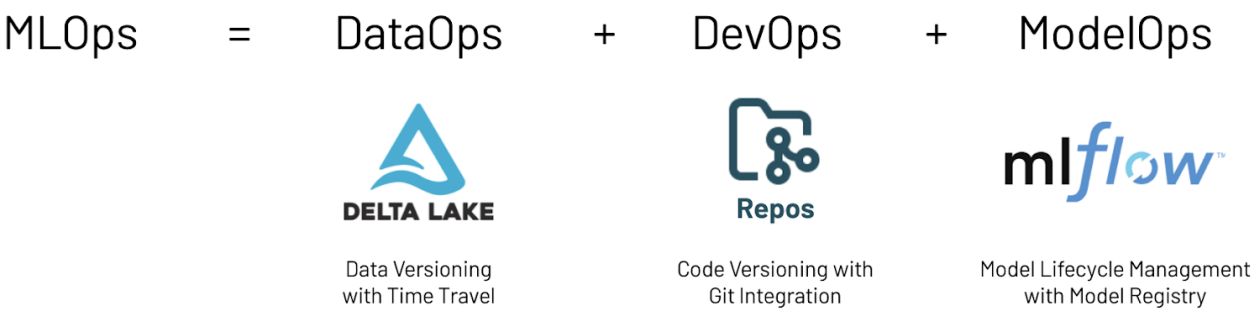
MLflow: is an open source platform to manage the end-to-end ML lifecycle.
It has components to monitor your model during training and running, reproducible runs, and sharing and deploying the models.
Load the model in production code and create pipeline.
Mlflow offers a set of lightweight APIs that can be used with any existing machine learning application or library (Tensorflow, PyTorch, XGBOost etc..), wherever you currently run ML code (e.g. in notebooks or the cloud).

Three distinct features are
MlFlow Tracking: You can define custom metrics so that after run you can compare the output to previous run using an user interface (UI).
MLFlow Projects allows you to create a pipeline
A code packaging format for reproducible runs using conda and docker, so you can share your ML code with others.
MLFlow models: A model packaging format and tools that let you easily deploy the same model on platforms such as docker, apache spark, Azure ML, and AWS SageMaker.

MLflow Model Registry: A centralized model store, set of APIs, and UI, to collaboratively manage the full lifecycle of MLflow models.

Databricks ML lifecycle:
Managing the end-to-end ML life cycle at scale from Databricks with Mlflow.
ML models are the result of compiling data and code into a machine learning model.

Databricks ML, easy to collaboration with team members and access controls on all types of objects (Notebooks, experiments, Models etc.).
It support multi language notebooks (support Python, SQL, R and Scala) within the same notebook.

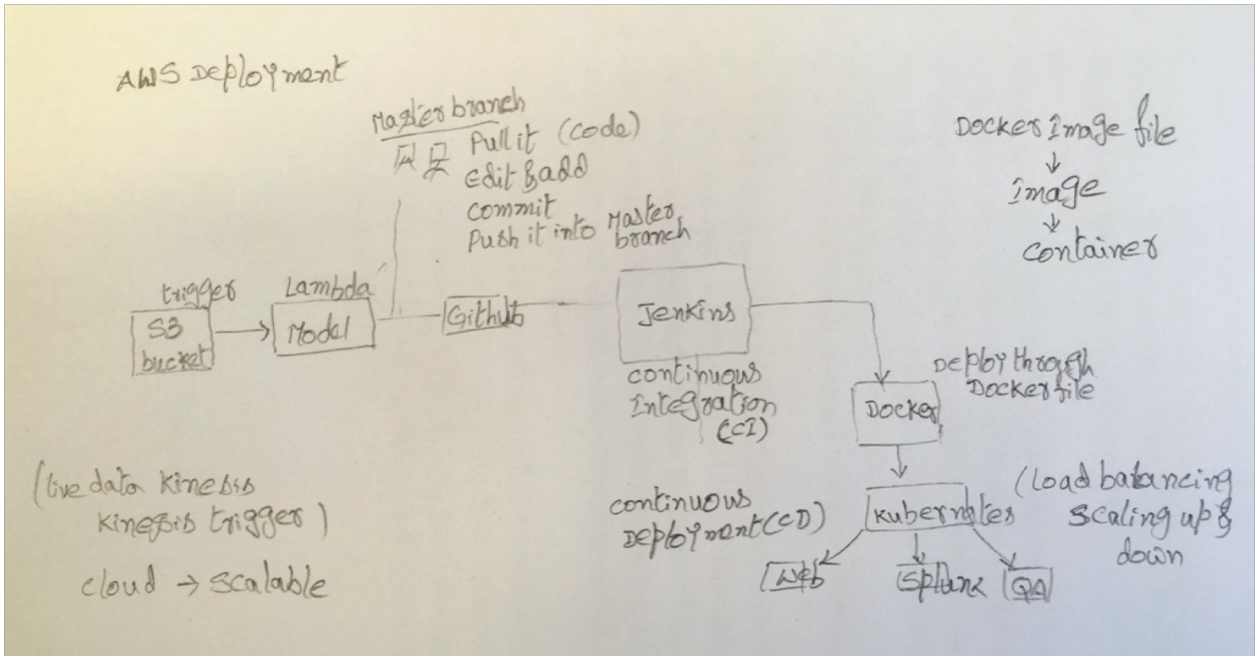


MLOPs = DataOps + DevOps + ModelOps
DataOps: Databricks ML is the only ML platform that provides built-in data versioning and governance. The exact version of the data is logged with every ML model that is trained on Databricks.

DevOps: Databricks ML provides integration with Git providers through its Repository feature, enabling data teams to follow best practice and integrate with CI/CD systems.

ModelOps: With managed Mlflow, Databricks ML provide a full set of features from tracking ML models with their associated parameter and metrics, to manage the deployment lifecycle, to deploying models on any platform (AWS, Azure, GCP).

Full reproducibility: providing a well-integrated solution for the full ML lifecycle means that work on Databricks ML is fully reproducible: data, parameters, metrics, models, code, compute configuration and library versions are all tracked and can be reproduced at anytime.


AWS deployment
Master branch
Pull it (code)
edit & add
commit
Push it into Master branch
Docker image file
image
container
trigger
Lambda
S3 bucket
Model
Github
Jenkins
continuous Integration (CI)
deploy through Docker file
Docker
(live data kinesis
kinesis triggers)
cloud → scalable
continuous deployment (CD)
kubernetes
(load balancing
scaling up & down
web
splunk
QA