**Content**

[1. What to learn 3](#_Toc21854580)

[2. AWS Provided Sample Questions 4](#_Toc21854581)

[1. TF-IDF (Term Frequency – Inverse Document Frequency) 4](#_Toc21854582)

[2.1.1. To learn more about it 4](#_Toc21854583)

[2. Fraud Detection 4](#_Toc21854584)

[3. Why Glue? 5](#_Toc21854585)

[4. Batch size and Learning rate in practice 5](#_Toc21854586)

[3. LinuxAcademy practice exam 6](#_Toc21854587)

[1. Difference between Kinesis Data Stream and Kinesis Firehose 6](#_Toc21854588)

[3.1.1. Kinesis Data Analytics 7](#_Toc21854589)

[3.1.2. Kinesis Firehose 7](#_Toc21854590)

[2. Characteristics of box plot 8](#_Toc21854591)

[3. Loss Functions (e.g. RMSE) 9](#_Toc21854592)

[4. Correlation 9](#_Toc21854593)

[5. CloudWatch Events – usecases 9](#_Toc21854594)

[6. SageMaker EI (Elastic Inference) 9](#_Toc21854595)

[4. Whizlabs – Free 15 questions 10](#_Toc21854596)

[1. targetAttributeName – data descriptors 10](#_Toc21854597)

[2. Hyperparameters 10](#_Toc21854598)

[3. Glue in more depth 10](#_Toc21854599)

[4. Some additional model evaluation methods 10](#_Toc21854600)

[5. KPI Charts 10](#_Toc21854601)

[6. Net Promoter Score (NPS) 10](#_Toc21854602)

[7. Data Imputation 11](#_Toc21854603)

[8. PR (Precision – Recall) Curve 11](#_Toc21854604)

[9. Binning 11](#_Toc21854605)

[5. AWS practice exam 12](#_Toc21854606)

[6. Quizlet and LinuxAcademy Flashcards 13](#_Toc21854607)

[1. SMOTE 13](#_Toc21854608)

[2. Apache Hive 13](#_Toc21854609)

[3. Learning Rate vs Batch Size 13](#_Toc21854610)

[4. Metrics 13](#_Toc21854611)

[6.4.1. Recall 14](#_Toc21854612)

[6.4.2. Precision 14](#_Toc21854613)

[5. Undersampling & Oversampling 14](#_Toc21854614)

[6. Threshold 14](#_Toc21854615)

[7. Others 16](#_Toc21854616)

[1. Collinearity 16](#_Toc21854617)

[2. Deploy custom algorithm using SageMaker 16](#_Toc21854618)

1. What to learn

Big data – kinesis and EMR.

1. learn 44 -> 35

2. whizlabs

3. quizlet

4. LA flashcards

5. LA flashcards

6. AWS Practice questions

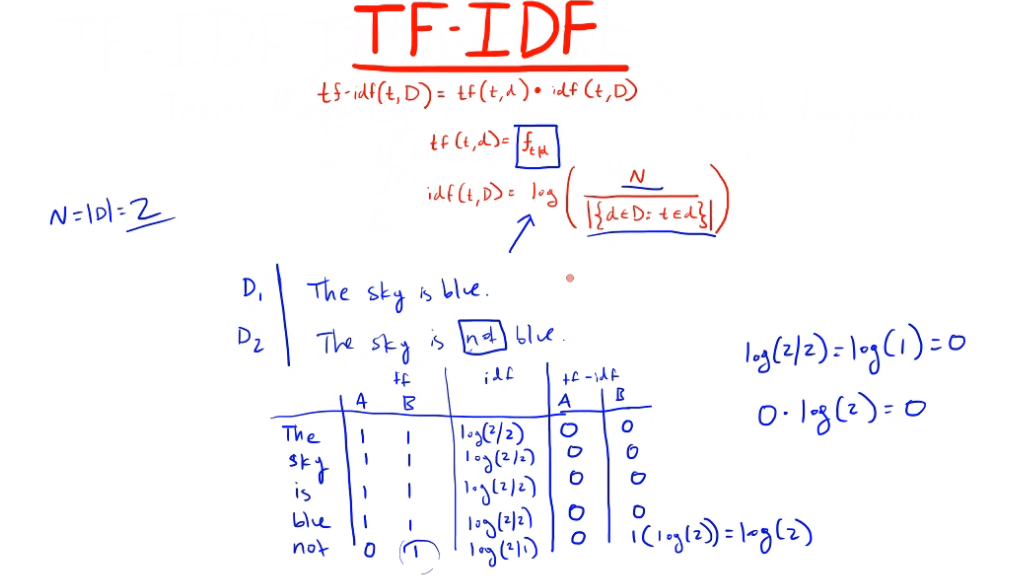
7. Practice Exam

8. Google Practice exams

9. Google ML course

1. AWS Provided Sample Questions
   1. TF-IDF (Term Frequency – Inverse Document Frequency)

The weight increases the more times a term occurs in the document, but that’s offset by the number of times the word appears in the entire dataset or corpus. This offset helps to remove the importance from really common words, like “the”, “or” “a”… Relevance ranking.



. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.

* + 1. To learn more about it
* Dampening
* magnitude
  1. Fraud Detection

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout.

* 1. Why Glue?

AWS Glue is the correct answer because this option requires the least amount of setup and maintenance since it is serverless.

* 1. Batch size and Learning rate in practice

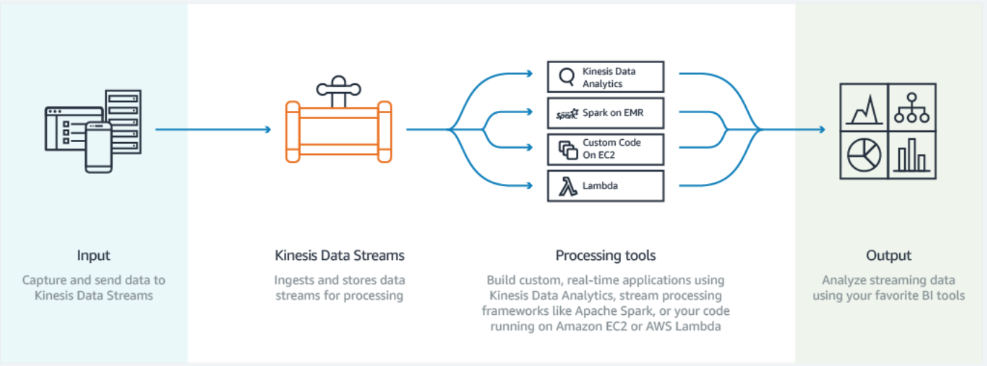
It is most likely that the **loss function is very curvy** and has **multiple local minima** where the training is getting stuck. **Decreasing** the **batch size** would help the Data Scientist stochastically **get out of the local minima saddles**. **Decreasing** the **learning rate** would **prevent overshooting the global loss function minimum**.

1. LinuxAcademy practice exam
   1. Difference between Kinesis Data Stream and Kinesis Firehose

*Firehose is send data to somewhere. Kinesis data streams on the other hand store data for up to 7 days for you to work with process.*

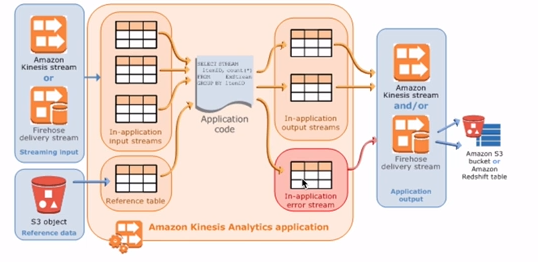
The main difference:

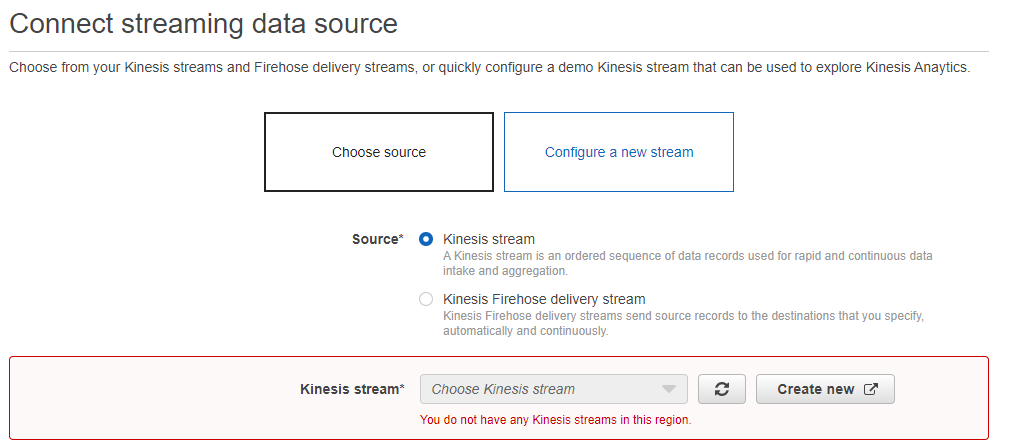
* Kinesis Streams just store the data in the stream. You should setup consumers to do something with it. So a lambda function and an EC2 instance can be reading from that stream at the same time and generating something or transforming data. Then an action can be taken, like send someone a message or store data in S3. Capture process and store data streams, **custom** application.
* Kinesis Firehose is meant to be used when you want to collect streaming data and store it in one of the 4 destinations. S3, Redshift, Splunk, Elasticsearch. Not used for real-time analytics. Can also transform data before delivering it. Load data streams into AWS data source. **Delivery stream.**





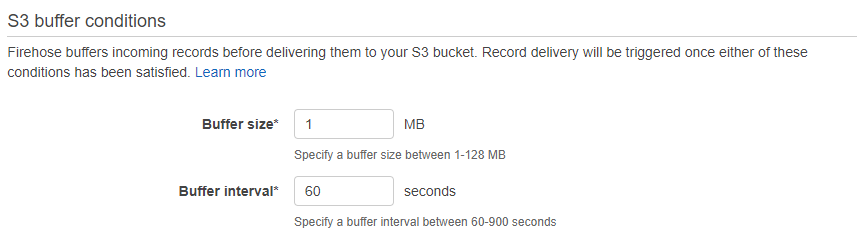
* + 1. Kinesis Data Analytics

Kinesis Data Analytics – Analyze data streams with SQS or Java. 



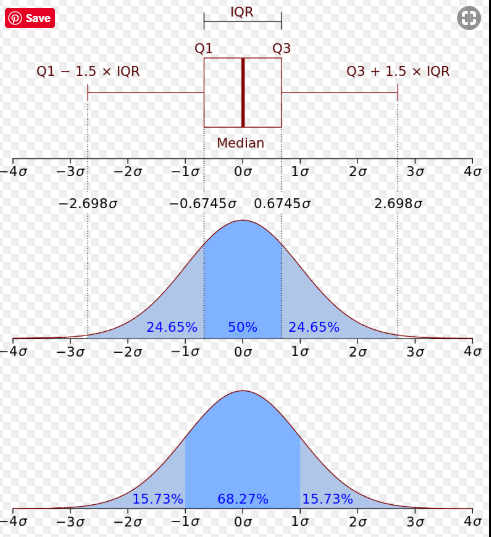
* + 1. Kinesis Firehose

As soon as one of these conditions become true, the buffer will be flushed and data will be written into S3. (it can be ES, RedShift or Splunk too).



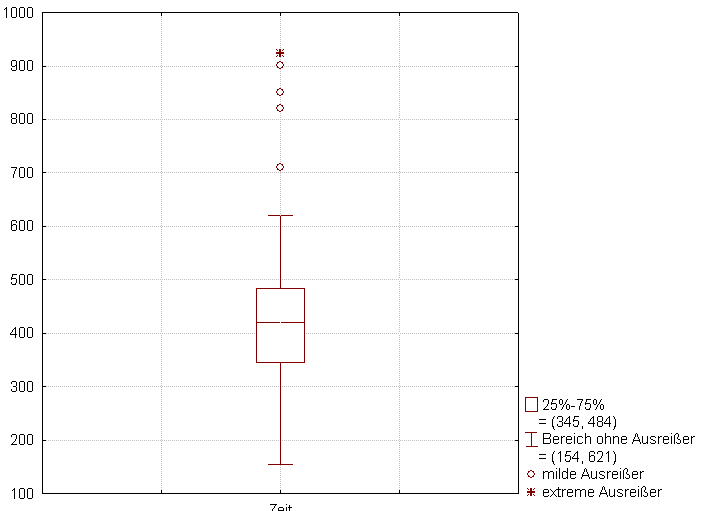
* 1. Characteristics of box plot

Boxplot and probability density function.



The points shown outside the ends of the whiskers are outliers or suspected outliers.

The ends of the lines (or whiskers) extending from the box indicate the minimum and maximum data values – unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.



* 1. Loss Functions (e.g. RMSE)

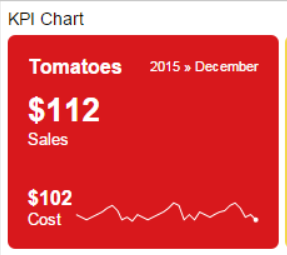
It's a method of evaluating how well your algorithm models your dataset. If your predictions are totally off, your loss function will output a higher number. If they're pretty good, it'll output a lower number. As you change pieces of your algorithm to try and improve your model, your loss function will tell you if you're getting anywhere.

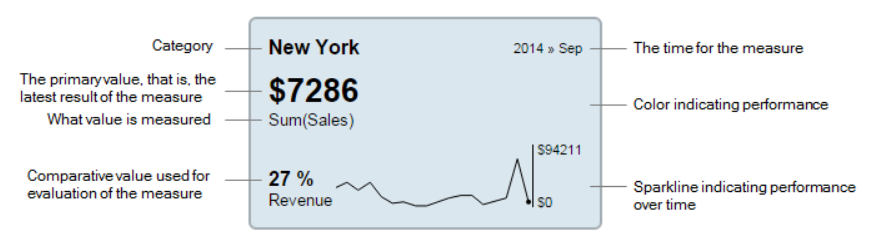
RMSE is a loss function for regression, but there are others for other types of problems.

* 1. Correlation
  2. CloudWatch Events – use cases
  3. SageMaker EI (Elastic Inference)

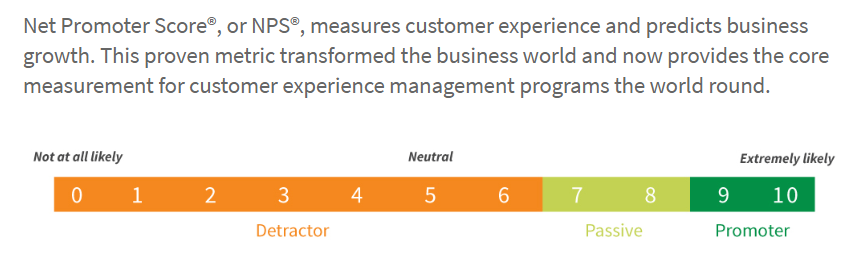
Elastic Inference (EI) is a resource you can attach to your EC2 CPU instances to accelerate your deep learning (DL) inference workloads. Amazon Elastic Inference allows you to attach GPU-powered acceleration to many amazon machine instances in order to reduce the cost of running deep learning inference. EI supports TF, MXNet, ONNX (Open Neural Network Exchange) through MXNet.

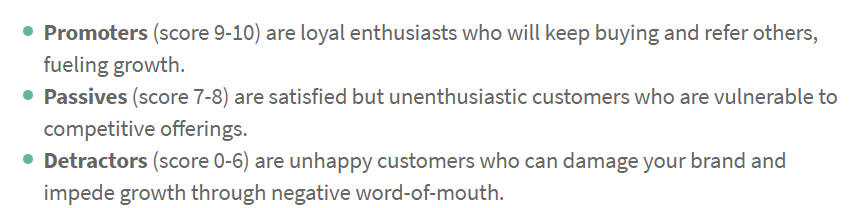
1. Whizlabs – Free 15 questions
   1. targetAttributeName – data descriptors
   2. Hyperparameters
   3. Glue in more depth
   4. Some additional model evaluation methods
   5. KPI Charts

The KPI chart is used to, at a quick glance, give information about the current performance of a company or organization. Factors, which are crucial for monitoring how the company performs, are measured and then presented in form of KPIs, Key Performance Indicators. The type of information that is shown varies. Examples of KPIs are information about net revenue, sales growth, and customer satisfaction.



* 1. Net Promoter Score (NPS)



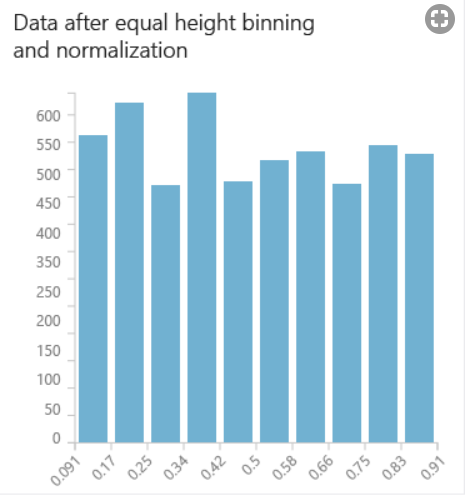


* 1. Data Imputation

How to replace the missing values in the dataset.

* 1. PR (Precision – Recall) Curve
  2. Binning

Binning or grouping data (sometimes called quantization) is an important tool in preparing numerical data for machine learning, and is useful in scenarios like these:

* A column of continuous numbers has too many unique values to model effectively, so you automatically or manually assign the values to groups, to create a smaller set of discrete ranges.
* For example, you could use entropy scores generated by Group Data into Bins to identify the optimal groupings of data values, and use those groups as features in your model.
* Replace a column of numbers with categorical values that represent specific ranges.
* For example, you might want to group values in an age column by specifying custom ranges, such as 1-15, 16-22, 23-30, and so forth for user demographics.
* A dataset has a few extreme values, all well outside the expected range, and these values have an outsized influence on the trained model. To mitigate the bias in the model, you might transform the data to a uniform distribution, using the quantiles (or equal-height) method.

1. AWS practice exam
2. Quizlet and LinuxAcademy Flashcards
   1. SMOTE

With datasets that are not fully populated, the Synthetic Minority Over-sampling Technique (SMOTE)

adds new information by adding synthetic data points to the minority class. This technique would be the

most effective in this scenario.

* 1. Apache Hive
  2. Learning Rate vs Batch Size
  3. Metrics

<https://www.mikulskibartosz.name/precision-vs-recall-explanation/>

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

Accuracy = TP + TN / Total

AUC - Area Under the ROC Curve (AUC) measures the ability of a binary ML model to predict a higher score for positive examples as compared to negative examples.

F1 - The macro-averaged F1-score is used to evaluate the predictive performance of multiclass ML models. This puts recall and precision together.

RMSE - The Root Mean Square Error (RMSE) is a metric used to evaluate the predictive performance of regression ML models.

Cut-off - ML models work by generating numeric prediction scores. By applying a cut-off value, the system converts these scores into 0 and 1 labels.

Accuracy - Accuracy measures the percentage of correct predictions.

Precision - Precision shows the percentage of actual positive instances (as opposed to false positives) among those instances that have been retrieved (those predicted to be positive). In other words, how many selected items are positive?

Recall - Recall shows the percentage of actual positives among the total number of relevant instances (actual positives). In other words, how many positive items are selected?

* + 1. Recall

TP / (TP + FN)

Recall answers What proportion of actual positives was identified correctly?

AWS: Recall shows the percentage of actual positives among the total number of relevant instances (actual positives). In other words, how many positive items are selected?

Think - If I was to recall what I did 10 years ago, I probably will guess something right, but there are other things that I will not guess because I forgot them. E.g. It is what I guessed right over (what i guess right + what I should have flagged as well)

* + 1. Precision

Precision answers What proportion of positive identifications was actually correct?

Precision = TP / (TP + FP)

AWS: Precision shows the percentage of actual positive instances (as opposed to false positives) among those instances that have been retrieved (those predicted to be positive). In other words, how many selected items are positive?

Think: Trying to be precise, but you ended up marking things as something that they aren't. E.g. What you guess right over (what you guess right + what we flagged, but we were wrong)

* 1. Undersampling & Oversampling

Resampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented)

* 1. Threshold

what about an email message with a prediction score of 0.6? In order to map a logistic regression value to a binary category, you must define a classification threshold (also called the decision threshold). A value above that threshold indicates "spam"; a value below indicates "not spam." It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.

* 1. Kinesis

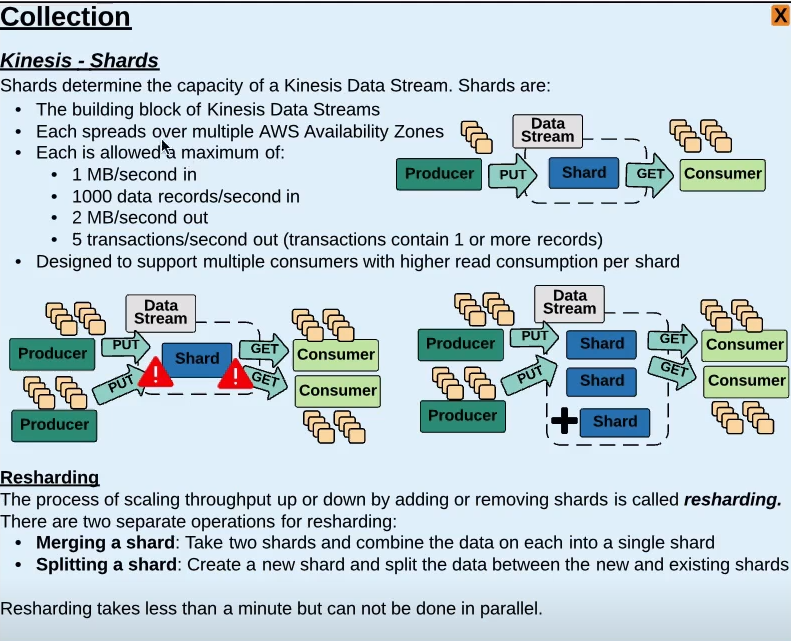
Kinesis Data Streams – more control over the scaling than Kinesis Firehose through shards.

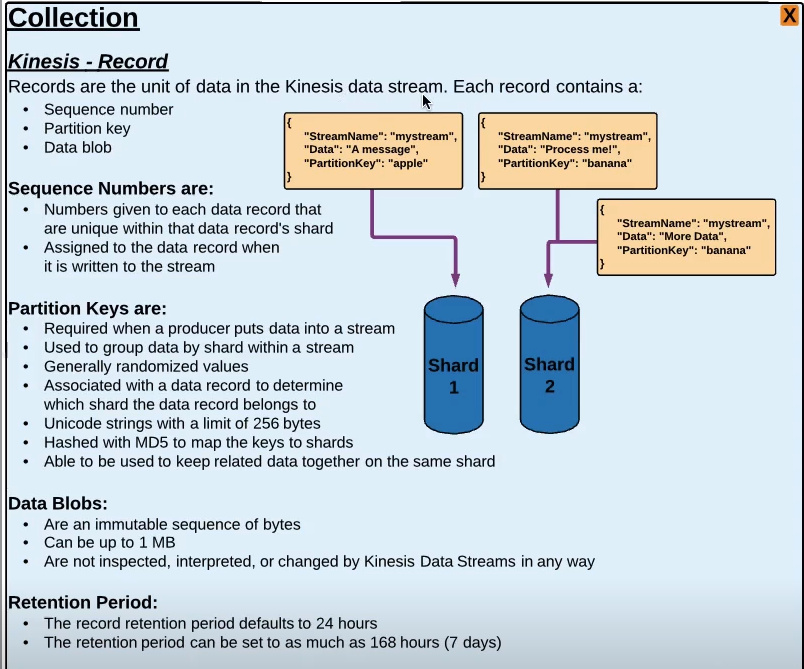
* + 1. Kinesis Data Streams



Real-time data streaming service that can continuously capture gigabytes of data per second from many sources like website clicks, financial transactions, and social media feeds…

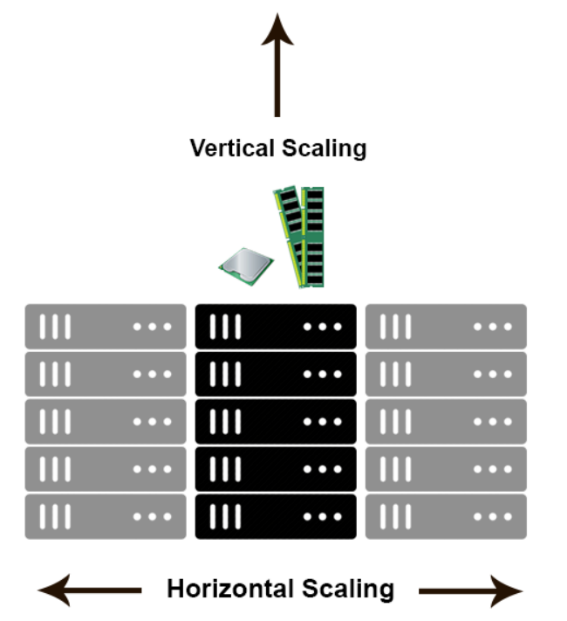
* + - 1. Kinesis Shards and Records





* + - 1. Sharding (database, general)

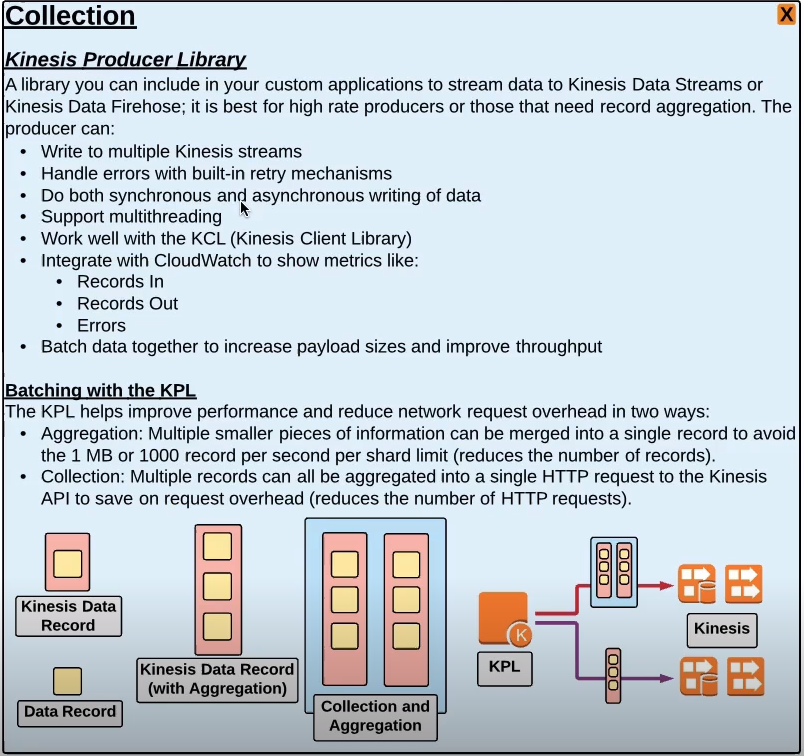
Sharding is a type of database partitioning that separates very large databases into smaller, faster, more easily managed parts called data shards. The word shard means a small part of a whole.

In the simplest sense, sharding your database involves breaking up your big database into many, much smaller databases that share nothing and can be spread across multiple servers.

Technically, sharding is a synonym for horizontal partitioning. In practice, the term is often used to refer to any database partitioning that is meant to make a very large database more manageable.

* + 1. Kinesis Producer Library

A library, you can include in your custom applications to steam data to Kinesis. It is best for high rate producers or those that need record aggregation.

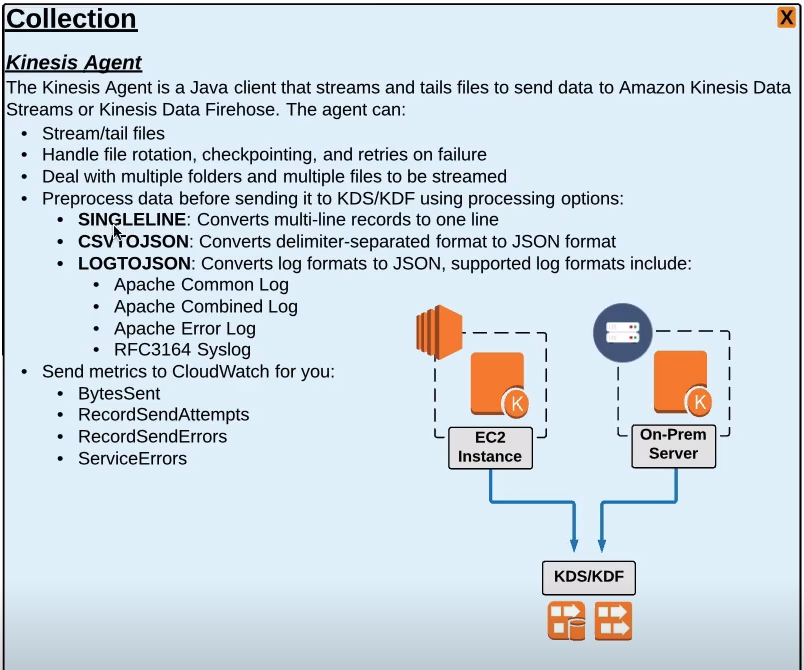


* + 1. Kinesis Client Library (KCL)



* + 1. Kinesis Agent

A java client that streams and tails files to send data into Kinesis Streams.



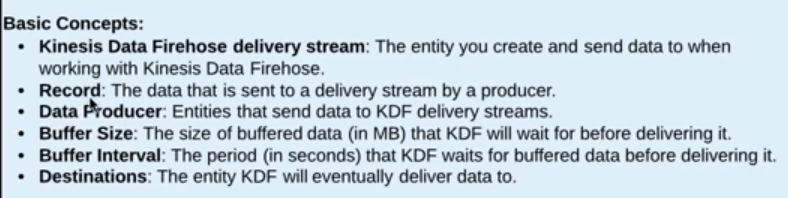
* + - 1. To tail files

It means tail -f command will wait for new strings in the file and show these strings dynamically. This command useful for observing log files.

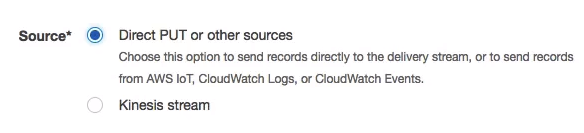
For example try, tail -f /var/log/messages.

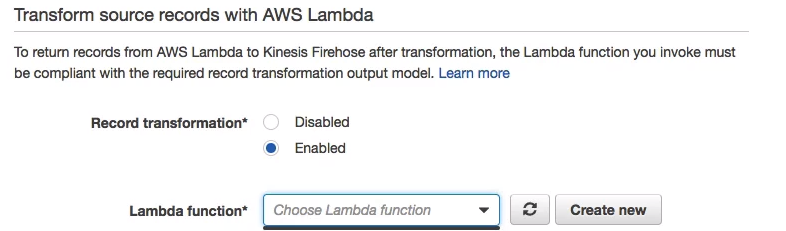
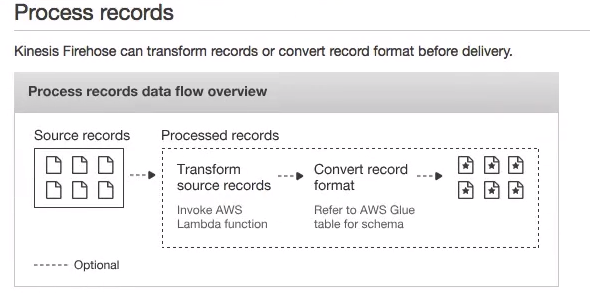
* + 1. Kinesis Data Firehose

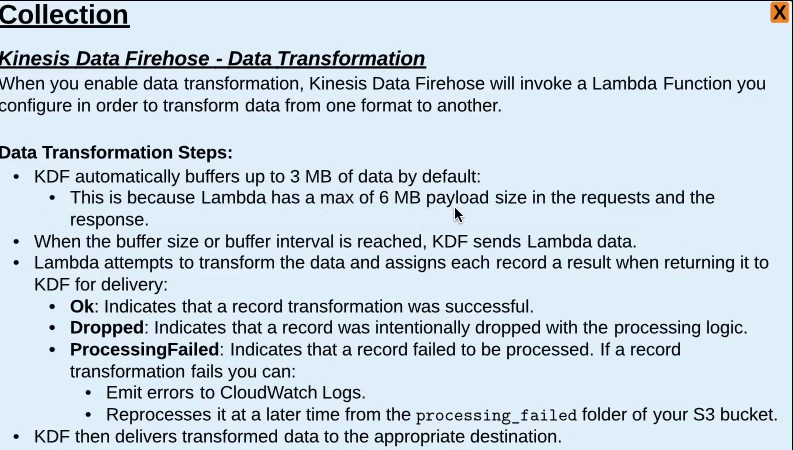
Concept of “Delivery Stream”

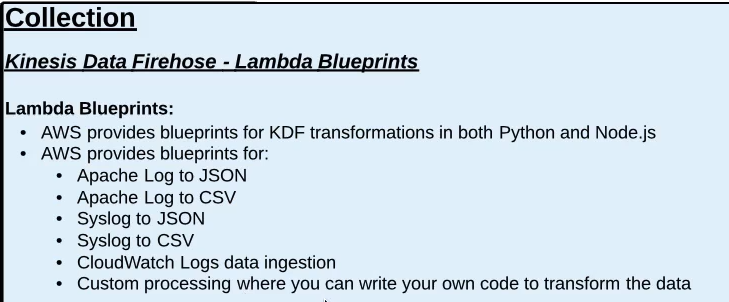


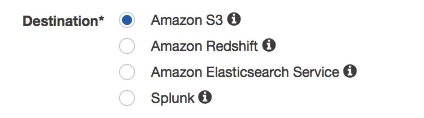
Buffer size – all the data will be queued up, but when it’s size reaches the buffer size it will be sent out to the destination. Similarly the buffer interval, when the interval gets reached …







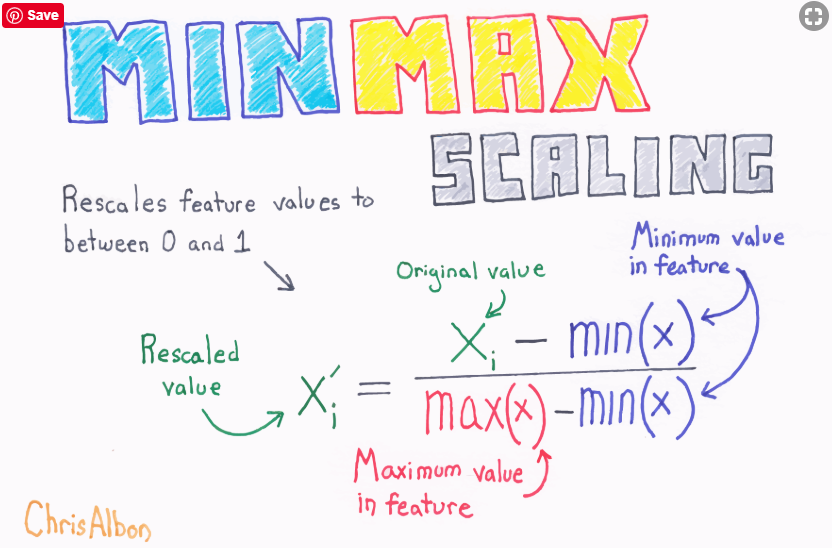


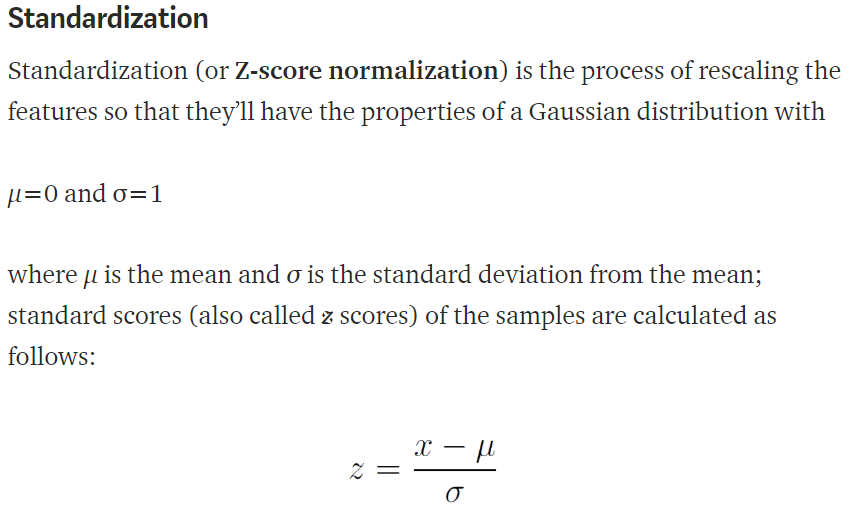


* 1. Orthogonal Sparse Bigram (OSB)
  2. Cartesian Product
  3. Normalization and Standardization

The two most discussed scaling methods are Normalization and Standardization. **Normalization** typically means rescales the values into a **range of [0,1]**. **Standardization** typically means **rescales** data to have a **mean of 0 and a standard deviation of 1** (unit variance).

**Normalization** is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. **It is required only when features have different ranges.**

For example, consider a data set containing two features, age(x1), and income(x2). Where age ranges from 0–100, while income ranges from 0–20,000 and higher. Income is about 1,000 times larger than age and ranges from 20,000–500,000. So, these **two features are in very different ranges.** When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically **influence** the result **more due to its larger value. But this doesn’t necessarily mean it is more important** as a predictor.



* 1. Cross Validation

There is always a need to validate the stability of your machine learning model. I mean you just can’t fit the model to your training data and hope it would accurately work for the real data it has never seen before. You need some kind of assurance that your model has got most of the patterns from the data correct, and its not picking up too much on the noise, or in other words its low on bias and variance.

The problem with classical evaluation techniques is that it does not give an indication of how well the learner will generalize to an independent/ unseen data set. Getting this idea about our model is known as Cross Validation.

* + 1. K-Fold Cross Validation
    2. Holdout Method
    3. Stratified K-Fold Cross Validation
    4. Leave-P-Out Cross Validation
  1. GPU FPGA ASIC CPU
  2. Sequential vs Randomized split

Sequential split: When you have something like time series

Randomized split: When you want an even random split

* 1. Backtesting

Backtesting is a term used in modeling to refer to testing a predictive model on historical data.

Basically, you use some of the historical data to check if a time series is working correctly.

Time-series data should be training and validated in order as it has, by definition, time as a potential influencing feature. A common method to validate time-series data is backtesting, or replaying the historical data as if it were new data, and then evaluating the model on how successful it predicted the historic values.

* 1. Gradient Descent
     1. Stochastic Gradient Descent
  2. Markov Decision Process

A Markov decision process (MDP) is a discrete time stochastic control process. It provides a mathematical framework for modeling decision making in situations where outcomes are **partly** **random** and **partly** **under** **the** **control** **of** **a** **decision** **maker**. MDPs are useful for studying optimization problems solved via dynamic programming and reinforcement learning. /wiki/

Has an Agent, in an Environment, with current state, actions it can perform, observation (what agent can do) with rewards. Episodes are iterations. Policy are the directions, the thing that is actually learning how to do the action or process.

**Reinforcement Learning.**

1. Others
   1. Collinearity

In geometry, collinearity of a set of points is the property of their lying on a single line. A set of points with this property is said to be collinear (sometimes spelled as colinear). In greater generality, the term has been used for aligned objects, that is, things being "in a line" or "in a row".

Collinearity, in statistics, correlation between predictor variables (or independent variables), such that they express a linear relationship in a regression model. When predictor variables in the same regression model are correlated, they cannot independently predict the value of the dependent variable. In other words, they explain some of the same variance in the dependent variable, which in turn reduces their statistical significance.

orthogonalization

* 1. Deploy custom algorithm using SageMaker
  2. Bias-Variance Tradeoff