**Content**

[1. What to learn 2](#_Toc21273435)

[2. AWS Provided Sample Questions 3](#_Toc21273436)

[1. TF-IDF (Term Frequency – Inverse Document Frequency) 3](#_Toc21273437)

[2.1.1. To learn more about it 3](#_Toc21273438)

[2. Fraud Detection 3](#_Toc21273439)

[3. Why Glue? 4](#_Toc21273440)

[4. Batch size and Learning rate in practice 4](#_Toc21273441)

[3. LinuxAcademy practice exam 5](#_Toc21273442)

[1. Difference between Kinesis Data Stream and Kinesis Firehose 5](#_Toc21273443)

[3.1.1. Kinesis Data Analytics 6](#_Toc21273444)

[3.1.2. Kinesis Firehose 6](#_Toc21273445)

[2. Characteristics of box plot 7](#_Toc21273446)

[3. RMSE 8](#_Toc21273447)

[4. Regularization 8](#_Toc21273448)

[5. Correlation 8](#_Toc21273449)

[6. Hyperparameter Tuning Strategies 8](#_Toc21273450)

[7. What to do when a model is underfit 8](#_Toc21273451)

[8. What to do when a model is overfit 8](#_Toc21273452)

[9. CloudWatch Events – usecases 8](#_Toc21273453)

[10. Inference pipelines 8](#_Toc21273454)

[11. Metrics 8](#_Toc21273455)

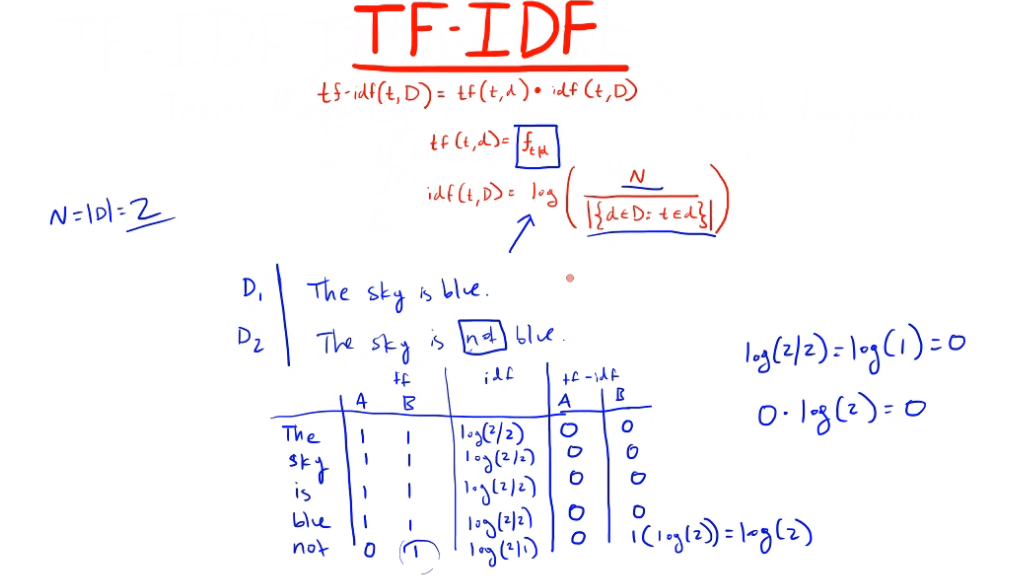
[4. AWS practice exam 9](#_Toc21273456)

1. What to learn

Big data – kinesis and EMR.

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

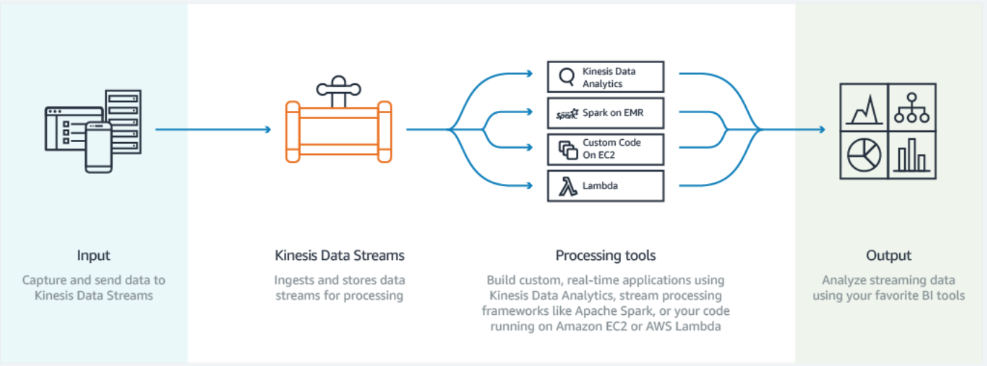
It is most likely that the **loss function is very curvy** and has **multiple local minima** where the training is getting stuck. **Decreasing** the **batch size** would help the Data Scientist stochastically **get out of the local minima saddles**. **Decreasing** the **learning rate** would **prevent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

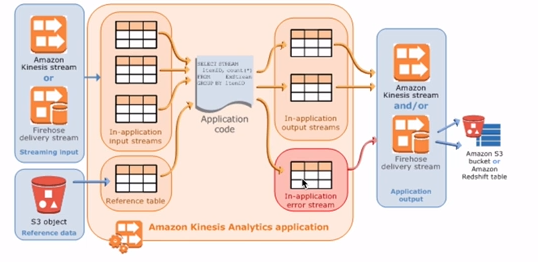
The main difference:

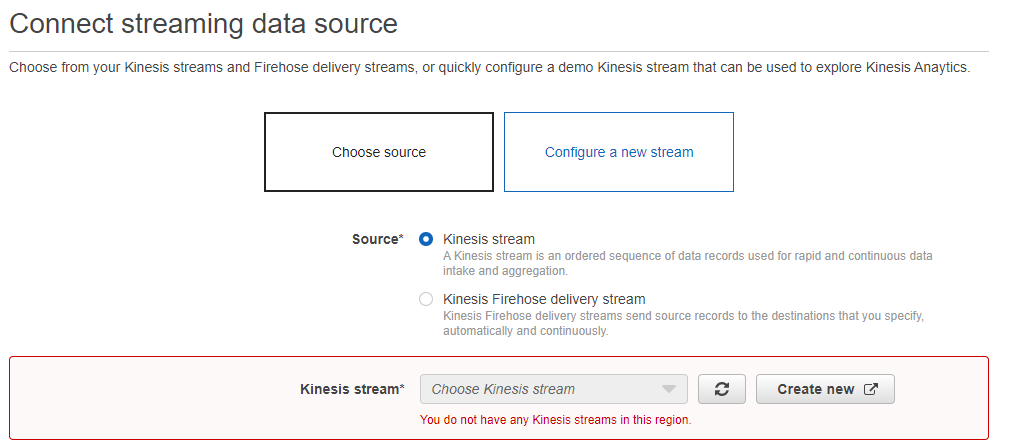
* Kinesis Streams just store the data in the stream. You should setup consumers to do something with it. So a lambda function and an EC2 instance can be reading from that stream at the same time and generating something or transforming data. Then an action can be taken, like send someone a message or 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.**





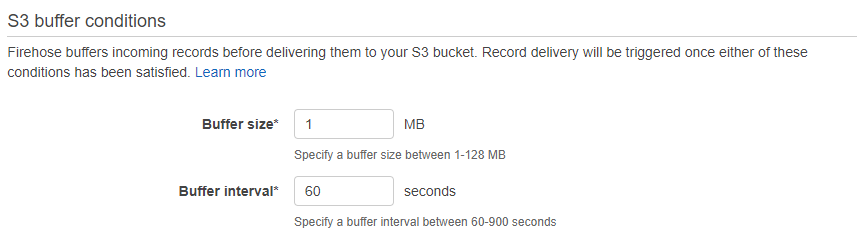
* + 1. Kinesis Data Analytics

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



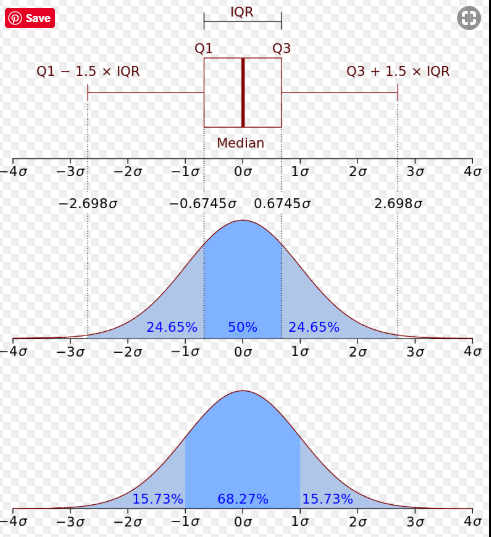
* + 1. Kinesis Firehose

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).



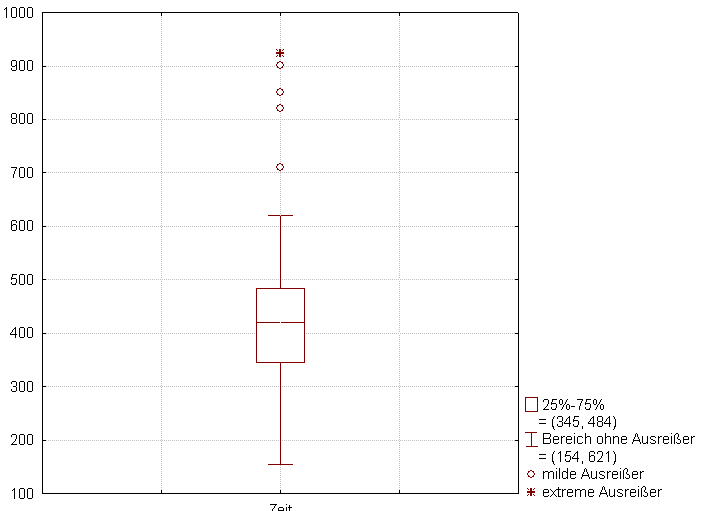
* 1. Characteristics of box plot

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. RMSE
  2. Regularization
  3. Correlation
  4. Hyperparameter Tuning Strategies
  5. What to do when a model is underfit
  6. What to do when a model is overfit
  7. CloudWatch Events – usecases
  8. Inference pipelines
  9. Metrics
* Recall (Sensitivity, TPR)
* Specificity (TNR)
* Precision
* Accuracy

1. Whizlabs – Free 15 questions
   1. targetAttributeName – data descriptors
   2. Hyperparameters
   3. Glue in more depth
   4. Some additional model evaluation methods
2. AWS practice exam