aws

Training Notes

AWS Certified Machine Learning Specialty

Karim El-Kobrossy
Neal Davis

DigitalCloud TRAINING

Table Of Contents

Introduction

These AWS cheat sheets are part of the **Ultimate Training Package for the AWS Certified Machine Learning Specialty** which include an on-demand video course and online practice exams to help you assess your exam readiness.

All of these training resources cover the latest MLS-C01 exam blueprint and are regularly updated.

The aim of putting this information together into one document is to provide a centralized list of facts you need to know before you sit your exam. This will shortcut your study time and provides the ideal opportunity for last-minute exam cramming.

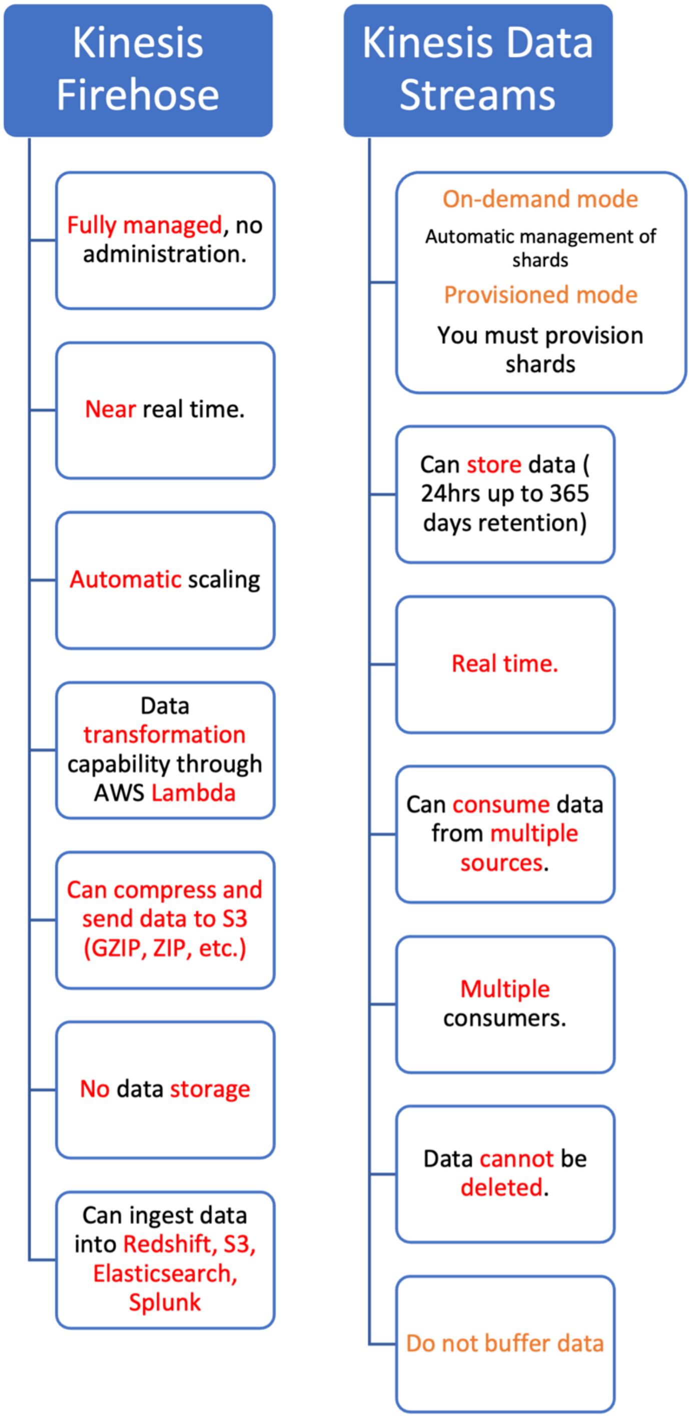
Take your studies offline with this downloadable PDF.

We wish you all the best for every step in your cloud journey!

Karim El-Kobrossy (Lead-Author)
& Neal Davis (Co-Author)

Data Engineering

Kinesis Firehose vs Kinesis Data Streams



Kinesis Firehose

- Fully managed, no administration.
- Near real time.
- Automatic scaling
- Data transformation capability through AWS Lambda
- Can compress and send data to S3 (GZIP, ZIP, etc.)
- No data storage
- Can ingest data into Redshift, S3, Elasticsearch, Splunk

Kinesis Data Streams

- On-demand mode

  Automatic management of shards

  Provisioned mode

  You must provision shards
- Can store data (24hrs up to 365 days retention)
- Real time.
- Can consume data from multiple sources.
- Multiple consumers.
- Data cannot be deleted.
- Do not buffer data

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data. It could be used for ETL purposes as well and it's a totally managed service in which we don't have access to the infrastructure

AWS Glue Crawler can identify data's schema while data is residing in the S3 bucket

AWS Glue data catalog is an index to the location, schema, and runtime metrics of your data

AWS Glue FindMatches ML can spot duplicates in the data

AWS Glue -> Ability to trigger ETL jobs based on a schedule or event

AWS Batch -> Fully managed service for running batch computing workloads not ETL

- Available as EventBridge (CloudWatch events) targets
- No managing resources

Amazon Redshift stores structured data while S3 stores structured and unstructured data

Amazon Kinesis Data Analytics

- Provides a function (RANDOM_CUT_FOREST) that can assign an anomaly score to each record based on values in the numeric columns.
- Enables developers to run SQL code against streaming sources and stream the results to an S3 bucket.
- Can spot "hotspots" using simple SQL commands

Amazon Kinesis Data Analytics could be used for:

- ETL
- Metric generation
- Responsive analytics

Amazon Kinesis video streams cannot stream data from multiple producers

Amazon Kinesis Firehose can transform the data from one format to another while Amazon Kinesis Data Streams cannot

If the capacity limits of Kinesis Data Streams are exceeded, the put data call will be rejected with a “ProvisionedThroughputExceeded” exception

Amazon Athena is a manged service which could be used to run SQL commands on data residing on S3

AWS DataSync is an online data transfer service that simplifies, automates, and accelerates moving data between on-premises storage systems and AWS storage services, and between AWS storage services

Amazon S3 is the storage service for amazon acting as a data lake. Most probably if a solution requires a data lake, then S3 is the answer

Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes

RDS and **Aurora** are relational database in which we should provision servers in advance

S3

- S3 Standard: Low retrieval time, high availability >=3 AZ.
- S3 Standard-IA: Low retrieval time, Low availability 1 AZ.
- S3 Intelligent tiering: for random access patterns, Low retrieval time, high availability >=3 AZ
- S3 Glacier deep archive: High retrieval time(hours-days), high availability >=3 AZ

DynamoDB: NoSQL data store, serverless

Elasticsearch: Indexing, clickstream analytics

S3 transfers data for training jobs while EFS doesn't transfer data

FSX for Lustre provides data to Sagemaker in a fast way for training different models

AWS Batch: runs batch jobs as docker images

AWS DMS: Continuous data replication, Source database remains available during migration, must create EC2 instance for replication job

AWS Step Functions: Visual workflow orchestration service

Exploratory Data Analysis

Amazon Athena can query data that sits in an S3 bucket provided that the schema is provided (AWS Glue crawler could automatically identify the schema)

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data

Missing data are one of the main problems which are inevitable in most of the datasets. There are many ways to overcome this problem such as:

- Deleting the missing rows entirely. While this is the easiest solution, it simply deletes many valuable information.
- Imputing missing data using mean. This is a bad technique to be used if the dataset contains outliers.
- Imputing missing data using mode. This is suitable for categorical features.
- Imputing missing data using KNN. This is suitable for numerical features.
- Imputing missing data using deep learning. This is suitable and accurate for categorical features.
- KNN is better suited for imputation of missing numerical values while deep learning is better suited for imputation of missing categorical values
- MICE technique is by far the most modern technique used for imputing missing data

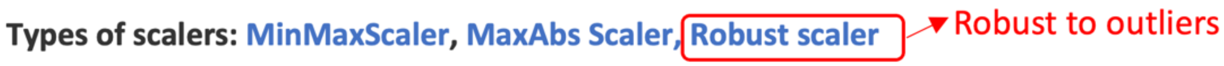
Exploratory Data Analysis

- Box and whisker: distribution of data + outliers
- Histogram: continuous distribution of data + frequency of given data + outliers
- Scatter matrix: observe correlations between pairs of features
- Heatmap
- Pair plot: correlation insights
- Scatter plot: plots data points of 2 features.
- "Tree map" visualize data in a hierarchical diagram
- Bubble chart: visualize 3D insights in 2D diagram
- Box and whisker and Histogram plots can spot outliers

While strong correlations have an impact on some algorithms, "decision trees" has feature selection embedded in it. So, strongly correlated features won't have a strong impact on decision trees.

A mapping technique where each label is mapped to a number representing its order between others is suitable for ordinal data.

PCA and T-SNE are dimensionality reduction techniques



Types of scalers: MinMaxScaler, MaxAbs Scaler, Robust scaler

Robust to outliers

One-hot encoding is best suitable for nominal data.

When the graph has a long right tail, it is right-skewed and vice versa. Log transform has the power of transforming right-skewed data to be more normally distributed.

SMOTE "Synthetic Minority Oversampling Technique" is a technique used to increase the number of the minority class by creating synthetic data points

Amazon Quicksight is an ML-powered business intelligence tool used to create visualizations and could be integrated with Amazon Athena

Amazon Quicksight ML insights can produce forecast on a given time-series data without machine learning prior knowledge

Amazon Ground Truth is a service with Amazon Sagemaker that labels datasets for further use in building machine learning models. Three options are available when using this service:

- Mechanical Turk: A team of global, on-demand workers from Amazon.
- Private labelling workforce: A team of workers from the company.
- Vendor: A selection of experienced vendors who specialize in providing data labelling services.

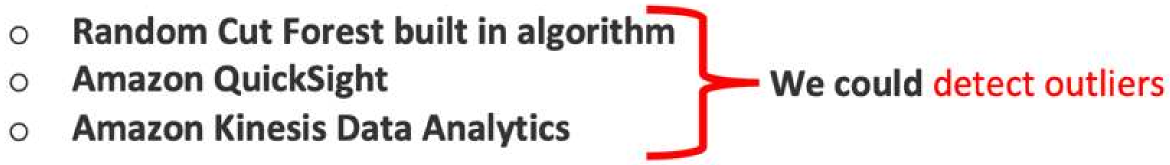
Bag of words: Matrix representations to describe the number of words within the text

TF-IDF: Gives more importance to words that are unique to the document

Word embedding: Words similar to each other are close to each other in this embedding space

Outliers

- Sometimes it is more appropriate to just remove outliers
- Sometimes you need outliers in your dataset as they resemble valuable information
- You could also apply transformation such as log transform to reduce the extreme variation



- ○ Random Cut Forest built in algorithm
- ○ Amazon QuickSight
- ○ Amazon Kinesis Data Analytics

We could detect outliers

Machine learning

To **avoid overfitting**, we could use

- Data augmentation
- Early stopping
- Regularization techniques
- Dropout
- Simplifying network's architecture

In **data augmentation**, the images are flipped, rotated, scaled, cropped, translated, etc. This will increase the number of images and will avoid overfitting the model as well.

When 2 features are found to have a strong correlation (positive or negative), one of them should be removed as they will affect the learning of the model whether it was a classification problem or a regression one. This is concerning feature-feature correlation not feature-target

We cannot simply train the model on high resolution, clear images and expect it to perform well on blurry images. This is a bad training example where we train the model on a distribution and test it on a completely different distribution.

High bias -> underfitting (too simple) (The model has not established the perfect relation between inputs and outputs)

High variance -> overfitting (too complex) (model not generalizing well to new data points)

Batch normalization overcomes vanishing and exploding gradient problems

RelU activation function does not have vanishing gradient problem

Adam, RMSProp and Adagrad and Stochastic gradient descent (SGD) are all **optimizers**, however SGD is the slowest among them.

Bagging: Generate new training sets by random sampling with replacement, each resampled model could be trained in parallel. Example: Random Forest.

Boosting: Training is sequential, each classifier considers the previous one. Example: AdaBoost, XGBoost.

Xavier and **normalized Xavier** are weight initialization techniques

PCA and **T-SNE** are dimensionality reduction techniques.

F1 is used to measure the overall performance of the model taking into consideration false positives and false negatives. **Recall** emphasizes on false negatives and **precision** emphasizes on false positives. **Accuracy** is better used when true positives and true negatives are more important.

Activation functions

- RelU
- Sigmoid
- Softmax
- Tanh

Softmax activation is used in the output layer when we want to predict only one class among other classes, so it makes a probability distribution among the outputs so that we can choose the highest probable output.

Sigmoid activation is used when we want to predict all classes present. The sigmoid outputs a probability for each class independent of one another.

L1 regularization is a regularization method which performs feature selection (some feature's coordinates can approach 0).

L2 regularization is a regularization method in which all features remain considered.

RMSE (Root Mean Squared Error) and **MSE** (Mean Squared Error) metrics and **MAE** (Mean Absolute Error) are used to evaluate regression models.

Accuracy, Precision, Recall, F1 score are used to evaluate classification models.

AUC (Area under curve) of the ROC curve is an evaluation of how well the model is performing. The higher the AUC, the better the model in distinguishing between classes

Small batch size tends not to get stuck in local minimum

A large batch size, however, results in a gradient with less stochasticity and optimization may *get stuck* in a local minimum or on a saddle point

Large learning rate can overshoot the correct solution.

Small learning rate slows training

Stratified K-fold cross validation technique is most suitable for unbalanced data to evaluate the model performance on unseen data

Recommender systems

- **Collaborative filtering** is used in recommender systems for calculating ratings based on ratings of similar users. Here, we have user A reading book X, if we got the books that other users recommend who read the same book X, then user A will most likely prefer those books. In fact, we do not know any features about other books, just that similar users liked them.
- Content-based filtering is used when user A read book X and we recommend him other books that have similar features as book X. Here, we know the features of other books, so we know that they could both be similar

You must shuffle your data. For example, the SGD algorithm is influenced by the order of the rows in the training data. Shuffling your training data results in better ML models because it helps the SGD algorithm avoid solutions that are optimal for the first type of data it sees, but not for the full range of data

Vanishing gradient problem could be mitigated by using:

- LSTM cells in a recurrent neural network
- Residual networks (ResNet)
- RelU activation function
- Batch normalization

Types of errors

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

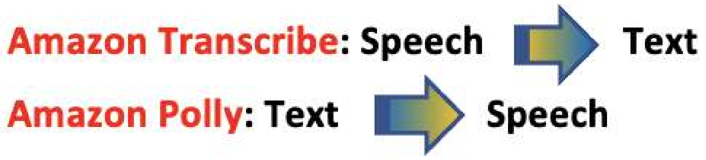
How could we prevent overfitting?

- Machine learning:
- L1 Regularization
- L2 Regularization
- Deep learning:
- Dropout
- Early stopping
- Image augmentation (more data)

AI services

Amazon Comprehend: Natural language processing service used to discover insights and relationships in text

Amazon Translate: Neural machine translation that offers language translation



Amazon Transcribe: Speech → Text

Amazon Polly: Text → Speech

Amazon Rekognition: Analyzes images and videos and extracts valuable information

AWS Deeplens: Deep learning-enabled video camera

Amazon Textract: Machine learning service that automatically extracts text, handwriting and data from scanned documents

Amazon Forecast: Fully managed forecasting service that uses statistical and machine learning algorithms to deliver forecasts

Amazon Lex: “Chatbot” -> Service for building conversational interfaces for applications using voice and text

Amazon Fraud Detector: Fully managed service that identifies potentially fraudulent online activities

Modelling

Elbow method is used in K-means algorithm in which it could determine a suitable number of clusters to segment the data into

Pipe mode is preferred over file mode due to:

- Shorter startup times because the data is being streamed instead of being downloaded to your training instances.
- Higher I/O throughputs due to our high-performance streaming agent.
- Virtually limitless data processing capacity.

Seq2Seq: machine translation, speech to text, etc.

Blazing Text

- Text classification: “supervised learning” in which it predicts labels for a sentence.
- Word2Vec: creates a vector representation of words. “CBOW”, “Skip-gram”, “Batch skip-gram”.

Object2Vec: creates low dimensional dense embeddings of high-dimensional objects.

Random Cut Forest: Anomaly detection.

KNN: Classification or regression.

IP insights: unsupervised learning in which it detects the anomaly score of a given IP address.

Factorization machines

- Sparse data
- RecordIO input not csv

AWS supports Bayesian, random search in hyperparameter tuning

K-means for segmentation, KNN for classification/regression.

In transfer learning we could Freeze all layers' weights except for the last couple layers including the output and re-train the model.

Use warm start to start a hyperparameter tuning job using one or more previous tuning jobs as a starting point

When training on more than one GPU, to ensure that you fully use each GPU, you must increase the mini-batch size linearly with each additional GPU

Training with multiple GPUs increase training speed if we increased both mini-batch size and learning rate

For **massive datasets** we can use Spark for preprocessing and Sagemaker for training, inference, tuning

Sagemaker notebooks are fully managed, so we don't have access to the underlying EC2 instance

Automatic model tuning

- Don't optimize too many hyperparameters at once.
- Limit your ranges to as small as possible.
- Use log scales when appropriate.
- Don't run too many parallel jobs.

A **VPC endpoint** enables connections between a virtual private cloud (VPC) and supported services, without requiring that you use an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Therefore, your VPC is not exposed to the public internet.

For training

- Use a built-in SageMaker algorithm
- Use prebuilt SageMaker container images with your own code (Script mode)
- Bring your own container image
- Extend a prebuilt container image

Use SageMaker Spark library to combine the power of Apache Spark in preprocessing and Sagemaker in training

Data parallelism

- Dataset is too big to fit into the memory!
- Gradients of small batches to be calculated on multiple GPUs

Model parallelism

- Model has too many layers/parameters and can't fit in a single GPU!
- Divide the model into pieces (5 layers for first GPU), (5 layers for second GPU)

When it comes to ensuring that an instance is highly available on the AWS cloud, more than one instance should be deployed across multiple availability zones.

XGBoost can take a csv file as an input provided that the headers had been removed

CloudWatch collects monitoring and operational data in the form of logs, metrics, and events.

Alarms are set using CloudWatch and can be further sent to an SNS topic to notify users.

Cloudtrail is used to audit activity to track user activity through API calls. It enables you to spot:

- Which users and accounts called AWS APIs for services that support CloudTrail.
- The source IP address the calls were made from.
- When the calls occurred.

AWS Autopilot currently only supports tabular data

SageMaker Debugger: SageMaker's feature that can

- Debug, monitor and profile training jobs in real time
- Detect non-converging conditions
- Optimize resource utilization by eliminating bottlenecks

Deployment

In **batch inferencing**, we get inferences on an entire dataset which is usually run on a schedule like every day or ever week. In **real-time inferencing** we are getting predictions on the spot.

Fargate could be used instead of EC2 instances when deploying containerized models as it can compute the capacity needed for inference and it is a managed service as well

SageMaker NeO for model's Optimization to be used on the cloud or edge devices. AWS IOT Greengrass can be used for deployment on edge devices

Amazon Elastic inference is used to apply the GPU power to the EC2 instance at much lower cost than using a GPU based instance (P class instances).

All models in Sagemaker are hosted in docker containers.

Inference pipeline: Linear sequence of 2-5 containers running sequentially at deployment. Each one of them is responsible for a task such as pre-processing, predictions, etc.

You can choose to restrict which traffic can access the internet by launching your Amazon SageMaker Studio and SageMaker notebook instances in a VPC

Amazon SageMaker supports automatic scaling (autoscaling) for your hosted models

Autoscaling dynamically adjusts the number of instances provisioned for a model in response to changes in your workload

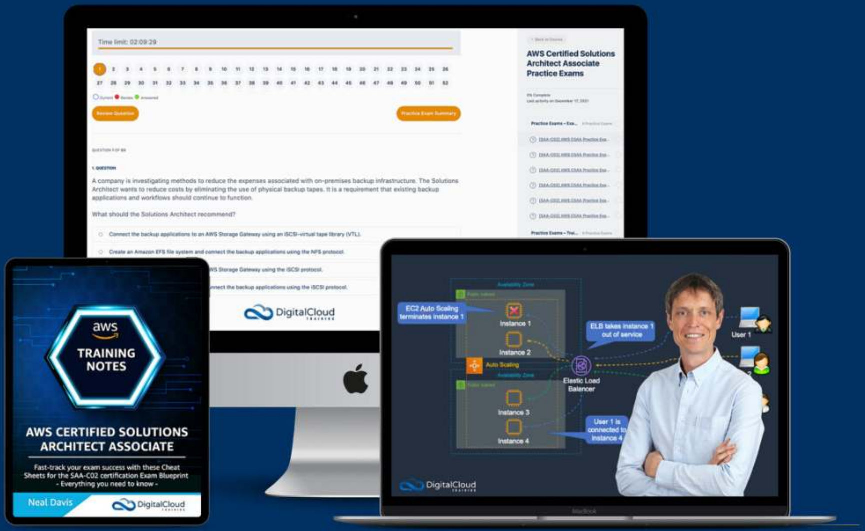
Model Monitor:

- Continuously monitors the quality of Amazon SageMaker machine learning models in production
- You can set alerts that notify you when there are deviations in the model quality
- **Types of monitoring**:
  - Monitor data quality
  - Monitor model quality
  - Monitor Bias Drift for Models in Production
  - Monitor Feature Attribution Drift for Models in Production

Elastic inference

- Network attached devices that work along with SageMaker instances in your endpoint to accelerate your inference calls
- Speed up the throughput and decrease the latency of getting real-time inferences at a fraction of the cost of using a GPU instance for your endpoint

EC2 Inf1 instances: Deliver high-performance ML inference at low cost


AWS Certified Solutions Architect Associate Practice Exams
aws
TRAINING NOTES
AWS CERTIFIED SOLUTIONS ARCHITECT ASSOCIATE
Fast-track your exam success with these Cheat Sheets for the SAA-C02 certification Exam Blueprint - Everything you need to know -
Neal Davis
DigitalCloud
EC2 Auto Scaling terminates instance 1
ELB takes instance 1 out of service
Instance 1
Instance 2
Instance 3
Instance 4
Auto Scaling
Elastic Load Balancer
User 1
User 1 is connected to instance 4

DigitalCloud
TRAINING

About Digital Cloud Training

Digital Cloud Training was created to help students achieve their career goals through high-quality AWS certification training resources. We provide a variety of certification training resources for Amazon Web Services (AWS) certifications that represent a higher quality standard than is otherwise available in the market.

Our popular AWS Certification exam preparation resources include instructor-led Video Courses, Hands-on Challenge Labs, in-depth Training Notes, Exam-Cram lessons for quick revision, Quizzes to test your knowledge and exam-difficulty Practice Exams to assess your exam readiness.

Join the AWS Community of over 750,000 happy students that are currently enrolled in Digital Cloud Training courses.

Visit digitalcloud.training for more information