Exploratory Data Analysis

Completed on 28-January-2021

Attempt	Marks Obtained	Your score	Time Taken	Result
01	0 / 8	0.0%	00 H 00 M 09 S	Failed

Domains wise Quiz Performance Report

No	Domain	Total Question	Correct	Incorrect	Unattempted	Marked as Review
1	Exploratory Data Analysis	8	0	1	7	0
Total	All Domain	8	0	1	7	0

Review the Answers

Sorting by All

Question 1 **Unattempted**

Domain :Exploratory Data Analysis

You work as a machine learning specialist for a mobile network operator who is building an analytics platform to analyze and optimize their operations by leveraging machine learning. You receive your data from source systems that send data in CSV format in real time. Your requirement is to transform the data to the parquet format before storing it on S3. From there you plan to use the data in SageMaker AutoPilot to help you find the best machine learning pipeline for your analytics problem.
Which option solves your data analysis machine learning problem in the most efficient manner?

A. Ingest the CSV data using MSK running on EC2 instances and use Kafka Connect S3 to convert the data to the parquet format.

B. Ingest the CSV data using Kinesis Data Streams. Convert the data to the parquet format using Glue.

C. Ingest the CSV data using Spark Structured Streaming in an EMR cluster. Convert data to the parquet format using Spark.

D. Ingest the CSV data using Kinesis Data Streams and use Kinesis Data Firehose, leveraging a Lambda function to transform the data from CSV to JSON, then convert the data to the parquet format. ✓

Explanation:

Answer: D

Option A is incorrect. Implementing the solution using MSK on EC2 instances requires more work than the other options.

Option B is incorrect. Unless you are using Glue streaming ETL, which is not explicitly stated in the question, you should not use Glue on streaming data.

Option C is incorrect. This option also requires more work (spinning up an EMR cluster) than simply using Kinesis Data Streams, Kinesis Data Firehose, and Lambda (all managed services).

Option D is CORRECT. You can ingest the streaming CSV data using Kinesis Data Streams, a managed service. Then use Kinesis Data Firehose and Lambda (also both managed services) to first convert the data from CSV to JSON (Kinesis Data Firehose transformation requires that the data be in the JSON format) then use the Kinesis Data Firehose parquet transformation to convert the data to parquet.

Reference:

Please see the **AWS blog** titled Stream Real-Time Data in Apache Parquet or ORC Format Using Amazon Kinesis Data Firehose.

Please refer to the **Kinesis Data Firehose developer guide** titled Converting Your Input Record Format in Kinesis Data Firehose.

Ask our Experts Rate this Question? ☺ ☹

View Queries open

Question 2 **Unattempted**

Domain :Exploratory Data Analysis

You are a machine learning specialist for a manufacturing company that has ingested structured and semi-structured manufacturing process data into their S3 buckets in their corporate data lake. Your data scientists now want to use SQL to run queries on this data to build manufacturing process KPI dashboards using a business intelligence tool.
Which option gives your data scientists the analysis and visualization capabilities they need in the most efficient manner?

A. **Transform the structured and semi-structured manufacturing process data into the parquet format using AWS Data Pipeline and then load the data into RDS from which your data scientists can run queries. Provide Kibana to your data scientists as the data visualization tool.**

B. **Catalog the structured and semi-structured manufacturing process data using a Glue crawler to populate your Glue data catalog. Then have your data scientists use Athena to run queries on their manufacturing data. Finally, the data scientists can build their KPI dashboards using the QuickSight Athena dataset feature.** ✓

C. **Transform the structured and semi-structured manufacturing process data then load the data into Aurora using an AWS Batch ETL job. Have your data scientists use a SQL tool to query the manufacturing data stored in Aurora and visualize the results by building their KPI dashboards using the QuickSight Aurora dataset feature**

D. **Transform the structured and semi-structured manufacturing process data into the parquet format using a Lambda function and use Kinesis Data Analytics to run queries and build the KPI dashboard visualizations.**

Explanation:

Answer: B

Option A is incorrect. Using AWS Data Pipeline to transform and load the data into RDS is not the most efficient option listed. Also, Kibana is best used as a visualization tool with AWS Elasticsearch, not RDS.

Option B is CORRECT. The AWS Glue crawler is the best option listed for making your manufacturing data available to a query tool like Athena by cataloging the data in your Glue data catalog. Athena is built to leverage the Glue data catalog to enable simple, efficient query capabilities for data stored in S3. Finally, QuickSight integrates directly with Athena through its Athena dataset connector. QuickSight has KPI dashboard capabilities built into it, making it the best BI visualization tool for your data scientists.

Option C is incorrect. Using AWS Aurora as the data store for your data scientist visualization work is far too complex. You would have to create the Aurora schema and database implementation. The Glue data catalog and Athena option is much more efficient.

Option D is incorrect. Using Lambda and Kinesis Data Analytics as the data source provider solution for your data scientist visualization work is far too complex. You would have to write the Lambda function code to process your manufacturing data. The AGlue data catalog and Athena option is much more efficient.

Reference:

Please see the **Towards Data Science article** titled Getting Started with Data Analysis on AWS.

Please refer to the **AWS Big Data blog** titled Analyzing Data in S3 using Amazon Athena.

Please review the **AWS Big Data blog** titled Build a Data Lake Foundation with AWS Glue and Amazon S3.

Ask our Experts Rate this Question?

View Queries open

Question 3 **Unattempted**

Domain :Exploratory Data Analysis

You are a machine learning specialist working for a language translation department of a major university. Your university has developed a mobile/web app that translates across different languages. You are now in the process of adding some of the more obscure languages in the far north area of the Arctic, such as Inuktun, Nganasan, and Dolgan. These languages are spoken by very few people in their regions so you have had to build your own data sources of the language patterns for each region.
Your machine learning team has decided to use Amazon Kendra to build an indexed searchable document repository. Your team needs to use the Kendra service to explore their language data in order to clean the data to prepare it for use in your language translation software. Your team has created your Kendra index and has added your data sources (HTML files, plain text files, PDFs, Word documents, PowerPoint presentations) in your S3 bucket to your index using the Kendra BatchPutDocument API call. However, you see in your CloudWatch logs an HTTP status code of 400 and some of your documents have not been successfully indexed.
What could be the source of the indexing failure?

A. **The total size of your files from your S3 bucket exceeds 25 MB**

B. **The text extracted from an individual Word document exceeds 5 MB**

C. **PDF documents are not supported by the Kendra BatchPutDocument API call**

D. **Microsoft PowerPoint presentations are not supported by the Kendra BatchPutDocument API call**

Explanation:

Answer: B

Option A is incorrect. The limit for the total size of your files from your S3 bucket is 50 MB, not 25 MB.

Option B is CORRECT. One of the limits for Kendra documents is that text extracted from an individual document cannot exceed 5 MB.

Option C is incorrect. Kendra supports the following unstructured document types HTML files, Microsoft PowerPoint presentations, Microsoft Word documents, plain text documents, and PDFs.

Option D is incorrect. Kendra supports the following unstructured document types HTML files, Microsoft PowerPoint presentations, Microsoft Word documents, plain text documents, and PDFs.

Reference:

Please see the **Amazon Kendra developer guide** titled Types of Documents.

Please refer to the **Amazon Kendra developer guide** titled Quotas for Amazon Kendra.

Please review the **Amazon Kendra developer guide** titled Common Errors.

Please refer to the **Amazon Kendra developer guide** titled BatchPutDocument.

Ask our Experts Rate this Question?

View Queries open

Question 4 **Unattempted**

Domain :Exploratory Data Analysis

You are a machine learning specialist at a company that is exploring conversational user interface application development. As an experiment, your team is building a natural language processing application. Your application needs to process the transcribed conversation data from your conversational user interface. For training, you are starting out with a dataset comprising 5 million sentences. You plan to run a model based on the Word2Vec algorithm to generate embeddings of the sentences. This will allow your team to make different types of

predictions.
Based on this example sentence: “My funy LARGE MEME went over the audiences head.”
Which operations should your team perform to sanitize and prepare the data in a repeatable manner? (CHOOSE THREE)

A. **Correct the spelling of "funy" to "funny" and “audiences” to “audience’s.”**

B. **Perform normalization by making the sentence lowercase.** ✓

C. **Using an English stopword dictionary, remove all stop words.** ✓

D. **Use One-hot encoding on the sentence.**

E. **Use part-of-speech tagging to keep the action verbs and the nouns only.**

F. **Perform tokenization of the sentence, creating a word vector.** ✓

Explanation:

Answers: B, C and F

Option A is incorrect. In natural language processing, the spelling of words has a relatively lower bearing on the importance of the word.

Option B is CORRECT. Using normalization you change all text so that it is on the same level. For example, converting all characters to lowercase. This allows algorithms like Bag Of Words and Word2Vec to perform more accurately.

Option C is CORRECT. Removing stop words like not, nor, never, etc, which are the most common words in a given language. Removing these words allows your algorithm to focus on differentiating words.

Option D is incorrect. One-hot-encoding is a technique used to encode categorical data.

Option E is incorrect. Using part-of-speech tagging to keep only the action verbs and nouns you would strip the conversation data of much of its meaning. The example sentence would become: Meme went head.

Option F is CORRECT. NLP algorithms like Word2Vec work best with tokenized data as their input.

Reference:

Please see the **article** titled Natural Language Processing: Text Data Vectorization.

Please refer to the **Towards Data Science article** titled NLP: Extracting the main topics from your dataset using LDA in minutes.

Please see the **Towards Data Science article** titled NLP Text Preprocessing: A Practical

Guide and Template.

Please see the **Towards Data Science article** titled 3 basic approaches in Bag of Words which are better than Word Embeddings.

Please see the **Towards Data Science article** titled Treat Negation Stopwords Differently According to Your NLP Task.

Please see the **Machine Learning Mastery article** titled Why One-Hot Encode Data in Machine Learning?.

Please see the **article** titled How to get started with Word2Vec — and then how to make it work.

Ask our Experts Rate this Question?

View Queries open

Question 5 **Unattempted**

Domain :Exploratory Data Analysis

You are a machine learning specialist at an online car retailer. Your machine learning team has been tasked with building models to predict car sales and customer conversion rates. The dataset you are using has a large number of features, over 1,000. Your team plans to use linear models, such as linear regression and logistic regression in a SageMaker Studio environment. When your team performs exploratory data analysis in their SageMaker Studio jupyter notebooks, they notice that many features are highly correlated with each other. Your tech lead has indicated that this may make your models unstable.
Which option would help you reduce the impact of having such a large number of features?

A. **Use dot product on the highly correlated features.**

B. **Use Principal Component Analysis (PCA) to create a new feature space** ✓

C. **One-hot-encode the highly correlated features.**

D. **Use TF-IDF encoding to reduce the impact of the highly correlated features.**

Explanation:

Answer: B

Option A is incorrect. Dot product, or matrix multiplication will not reduce the impact of

having 1,000 features in your dataset. It is used in deep learning for operations such as the Softmax function.

Option B is CORRECT. Principal Component Analysis (PCA) is a very common technique used in machine learning to reduce the dimensionality of your dataset. Reducing the dimensionality reduces the impact of having a large number of correlated features.

Option C is incorrect. One-hot-encoding is a technique used to encode categorical data. One-hot-encoding will actually increase the number of features in your dataset.

Option D is incorrect. TF-IDF, or Term Frequency Inverse Document Frequency, is used to indicate the importance of a word in a document in a collection or corpus. You are dealing with sales data and conversion rates, not text datasets.

Reference:

Please see the **Amazon SageMaker developer guide** titled Amazon SageMaker Studio.

Please refer to the **Data Science Bootcamp article** titled Understand Dot Products Matrix Multiplications Usage in Deep Learning in Minutes — beginner friendly tutorial.

Please see the **Data Science Bootcamp article** titled Understand the Softmax Function in Minutes.

Please see the **article** titled A simple guide to One-hot Encoding, tf and tf-idf Representation

Ask our Experts Rate this Question?

View Queries open

Question 6 **Unattempted**

Domain :Exploratory Data Analysis

You work on a machine learning team at an online reseller of consumer products. You are performing feature engineering of your product data where you have a large, multi-column dataset with one column missing 40% of its data. Your team lead thinks that you can use some of the columns in the dataset to create the missing data.
Which feature engineering is the best approach to use to create approximate replacements for the missing data while also preserving the integrity of the dataset?

A. **Binning**

B. **Yeo-Johnson transformation**

C. **Multivariate imputation** ✓

D. **Mean imputation**

Explanation:

Answer: C

Option A is incorrect. Binning is used for grouping values. It is used to minimize the impact of observation errors. Binning would not help you create approximate replacements for missing values

Option B is incorrect. Yeo-Johnson transformation is used to give your data a more gaussian distribution. It is not used to create approximate replacements for the missing data.

Option C is CORRECT. With multivariate imputation you use other variables in the data set to predict missing values. This

Option D is incorrect. Mean imputation replaces the missing values with the mean of observed values of that variable. This approach is the most simplistic method of imputation of missing values. The multivariate imputation method is much more accurate.

Reference:

Please see the **article** titled Binning in Data Mining.

Please refer to the **Machine Learning Mastery article** titled How to Use Power Transforms for Machine Learning.

Please see the **article** titled Multiple Imputation in a Nutshell

Ask our Experts Rate this Question? ☺ ☹

View Queries open

Question 7 **Unattempted**

Domain :Exploratory Data Analysis

You are a machine learning specialist at a financial services company. Your team has recently been assigned a project to prepare financial risk data and use it in a risk management machine learning model. The project is on an expedited schedule so you need to produce your engineered data as quickly as possible.

Which AWS service(s) will allow you to engineer your risk data as expeditiously as possible?

A. **SageMaker Studio**

B. **SageMaker Augmented AI**

C. **Deep Learning Containers**

D. **SageMaker Processing** ✓

Explanation:

Answer: D

Option A is incorrect. You could use SageMaker Studio to perform your data engineering tasks, but more of the infrastructure and coding work would have to be done by you and your team when compared to using SageMaker Processing.

Option B is incorrect. SageMaker Augmented AI is used to leverage human review of low confidence predictions. It wouldn't help your team expedite your data engineering work.

Option C is incorrect. Deep Learning Containers are a set of Docker images used for training and serving models in TensorFlow, PyTorch, and Apache MXNet. Deep Learning Containers wouldn't help your team expedite your data engineering work.

Option D is CORRECT. SageMaker Processing is an AWS managed service that you can use to run data engineering workloads in SageMaker using a simple SageMaker Processing APIs. SageMaker Processing manages your SageMaker environment for you in a processing container. This managed service removes much of the infrastructure and coding work need to perform data engineering tasks.

Reference:

Please see the **Amazon SageMaker developer guide** titled Process Data and Evaluate Models.

Please see the **Amazon SageMaker developer guide** titled Using Amazon Augmented AI for Human Review.

Please see the **Amazon SageMaker developer guide** titled Amazon SageMaker Studio.

Please see the **GitHub repository** titled Amazon SageMaker Processing jobs.

Please see the **AWS Deep Learning Containers development guide** titled What are AWS Deep Learning Containers?

Ask our Experts Rate this Question?

View Queries open

Question 8 **Incorrect**

Domain :Exploratory Data Analysis

You are a machine learning specialist at a security firm that is building a video surveillance service to be used by police departments across the country. This service needs to process the frames of streaming video to find suspicious activity in public places such as train stations, subway platforms, etc. In order to accomplish this task your team needs to use a machine learning technique to find objects in the video frames that are on a list of objects identified as being potentially dangerous, such as weapons. Your requirement is to label your images by identifying the contents of your images at the pixel level for high accuracy.
Which AWS service gives you the labeling accuracy your project requires?

A. SageMaker Ground Truth Bounding Box labeling task

✓ B. SageMaker Ground Truth Image Classification labeling task

C. SageMaker Ground Truth Image Semantic Segmentation labeling task

D. SageMaker Ground Truth Named Entity Recognition labeling task

Explanation:

Answer: C

Option A is incorrect. Using the SageMaker Ground Truth Bounding Box labeling task, you can identify the pixel location of an object, but not identify the contents of an image at the pixel level.

Option B is incorrect. Using the SageMaker Ground Truth Image Classification labeling task, your workers will classify your images using a predefined set of labels that you specify, but not identify the contents of an image at the pixel level.

Option C is CORRECT. Using the SageMaker Ground Truth Image Semantic Segmentation labeling task your workers classify pixels in the image into a set of predefined labels or classes. This will give you the pixel level label identification accuracy you require.

Option D is incorrect. The SageMaker Ground Truth Named Entity Recognition labeling task is used to extract information from unstructured text and classify it into predefined categories.

Reference:

Please see the **Amazon SageMaker developer guide** titled Use Amazon SageMaker Ground Truth to Label Data.

Please see the **Amazon SageMaker developer guide** titled Bounding Box.

Please see the **Amazon SageMaker developer guide** titled Image Semantic Segmentation.

Please see the **Amazon SageMaker developer guide** titled Image Classification (Single Label).

Please see the **Amazon SageMaker developer guide** titled Named Entity Recognition

Ask our Experts

Rate this Question?

View Queries open

Finish Review

Certification

Cloud Certification

Java Certification

PM Certification

Big Data Certification

Company

Become Our Instructor

Support

Discussions

Blog

Business

Support

Contact Us

Help Topics

Join us on Slack!

Join our open **Slack community** and get your queries answered instantly! Our experts are online to answer your questions!

Follow us