← Back **Graded Quiz: Data Engineering for Machine Learning using Apache Spark**
Graded Assignment • 30 min

English ∨ **Due** Nov 13, 11:59 PM IST

Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 70%. We keep your highest score.

Next item →

1. Which API does Apache Spark Structured Streaming use for processing streaming data?

1 / 1 point

○ Spark Streaming API

○ Kafka Streaming API

○ Spark SQL API

● DataFrame and Dataset APIs

✓ **Correct**
Correct! Apache Spark Structured Streaming uses the DataFrame and Dataset APIs for processing streaming data. These APIs provide a consistent programming model with batch processing in Spark SQL.

2. Which output mode in Spark Structured Streaming is particularly useful for managing late-arriving data points?

1 / 1 point

● Update mode

○ Append mode

○ Overwrite mode

○ Complete mode

✓ **Correct**
Correct! The update output mode in Spark Structured Streaming is especially helpful for enabling and managing late-arriving data points. It allows updating existing rows in the output.

3. Which Spark feature allows you to query a dataframe?

1 / 1 point

○ Pipeline

○ RDD

● SparkSQL

○ DataFrame

✓ **Correct**
Correct! SparkSQL provides a high-level interface for querying structured and semi-structured data in Spark using SQL-like syntax.

4. Which transformer/extractor counts the occurrences of each term in the text and constructs a vector representation?

1 / 1 point

○ StopWordsRemover

● CountVectorizer

○ StringIndexer

○ StandardScaler

✓ **Correct**
Correct! CountVectorizer is the transformer/extractor that counts the occurrences of each term in the text and constructs a vector representation.

5. In a certain SparkML pipeline, there are these 3 stages StandardScaler, VectorAssembler, and Linear Regression. Which of the following is the correct order?

1 / 1 point

○ StandardScaler, VectorAssembler, Linear Regression

○ Linear Regression, StandardScaler, VectorAssembler

● VectorAssembler, StandardScaler, Linear Regression

○ StandardScaler, Linear Regression, VectorAssembler

✓ **Correct**
Correct! This is the correct order. In the pipeline, the vectors are assembled using VectorAssembler, then the assembled vectors are scaled using StandardScaler, and finally, Linear Regression is applied.

6. In which phase of the ETL are you likely to encounter "save data to parquet file"?

1 / 1 point

● Load

○ Extract

○ Transform

○ Model Building

✓ **Correct**
Correct! In the Load phase of the ETL process, you are likely to encounter "save data to a Parquet file." This phase involves loading the transformed data into the target system, which can include saving the data to different file formats, such as Parquet.

7. What does a StringIndexer do? **1 / 1 point**

● Converts categorical string columns into numerical indices.

○ Converts numerical columns into strings.

○ Indexes strings so that they can be accessed quicker.

○ Converts floating point columns into string indices.

✓ **Correct**
Correct! The StringIndexer is used to convert categorical string columns into numerical indices.

8. What does a Tokenizer do? **1 / 1 point**

● Converts text into words.

○ Converts text into symbols called tokens.

○ Assigns a token to each row of the dataset.

○ Is used to compress text.

✓ **Correct**
Correct! A Tokenizer is used to convert text into words. It takes a sentence or a document as input and breaks it down into individual words or tokens.

9. Which Pyspark component is used to load a stored model? **1 / 1 point**

● PipelineModel

○ Pipeline

○ Model

○ ModelLoader

✓ **Correct**
Correct! The PipelineModel component in PySpark is used to load a stored model. It allows you to load and use a previously trained pipeline model for making predictions on new data.

10. In which phase of the ETL process would you typically perform "data validation"? **1 / 1 point**

○ Load

○ Extract

○ Model Building

● Transform

✓ **Correct**
Correct! In the Transform phase of the ETL process, you would typically perform "data validation." This phase involves ensuring the quality and integrity of the extracted data, such as checking for missing values, data type inconsistencies, or any other data quality issues before further processing.