Hide menu

Spark for Data Engineering

- **Reading:** Course Introduction — 2 min
- **Video:** Spark Structured Streaming — 4 min
- **Video:** GraphFrames on Apache Spark — 5 min
- **Video:** ETL Workloads — 4 min
- **Video:** Introduction to the pipeline editor in Elyra (Optional) — 11 min
- **Ungraded Plugin:** Reading: Create component oriented data science pipelines using CLAIMED, Elyra, KubeFlow Pipelines, MLX and Kubernetes — 10 min
- **Ungraded App Item:** Hands-on Lab: ETL using Apache Spark — 20 min
- **Reading:** Summary & Highlights — 3 min
- **Practice Quiz:** Practice Quiz: Spark for Data Engineering — 6 questions
- **Quiz:** Graded Quiz: Spark for Data Engineering — 10 questions

Graded Quiz: Spark for Data Engineering

Quiz • 20 min

Submit your assignment

Due Aug 6, 11:59 PM EDT **Attempts** 3 every 8 hours

Receive grade

To Pass 60% or higher

Like Dislike Report an issue

Start assignment

Your grade

–

1. Select the option where all four statements about streaming data characteristics are correct. **1 point**

- ○ Data is generated in finite, small batches; often originates from more than one source; is often available as a complete data set; requires incremental processing .
- ◉ Data is generated continuously; often originates from more than one source; is unavailable as a complete data set; requires incremental processing.
- ○ Data is generated incrementally; often originates from more than one source; is unavailable as a complete data set; requires incremental processing.
- ○ Data is generated incrementally; often originates from more than one source; is unavailable as a complete data set; requires batch processing.

2. Select the data sink option that is **not** fault-tolerant and that is recommended for debugging only. **1 point**

- ◉ Console and Memory
- ○ Foreach and ForeachBatch
- ○ Kafka
- ○ Files

3. Select the answer with the options that best completes the following statement: **1 point**

Apache Spark Structured Streaming processes a data stream with the Spark SQL engine _____.

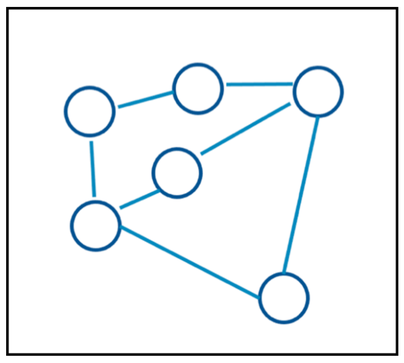
- ○ Extended SQL APIs
- ○ RDD APIs
- ○ Structured Streaming specific APIs
- ◉ Dataset and DataFrame APIs

4. Select the website where you can find and download the GraphFrames package. **1 point**

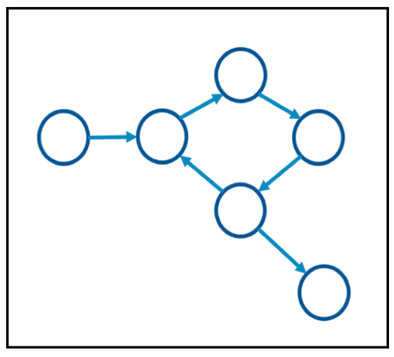
- ○ On the GraphFrames.com website
- ○ On the sparkpackages.org website
- ◉ On the spark-packages.org website.
- ○ On the Spark.com website

5. Identify which options correctly describe a directed graph and an undirected graph. (Multiple answers) **1 point**

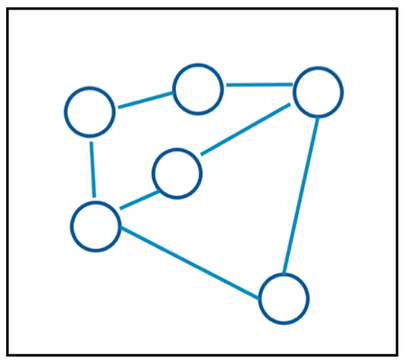
☐ A directed graph contains edges with a single direction between two vertices, indicating a one-way relationship, illustrated using lines **without** arrows.



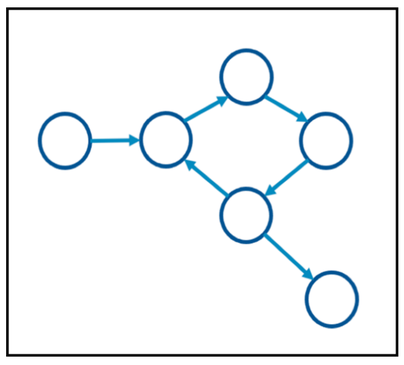
☑ A directed graph contains edges with a single direction between two vertices, indicating a one-way relationship, illustrated using lines **with** arrows.



☑ Undirected graphs have edges representing a relationship **without** a direction, illustrated using lines without arrows.



☐ Undirected graphs have edges representing a relationship without a direction, illustrated using lines **with** arrows.



6. Select the option that lists the correct order of these ETL workflow items. **1 point**

Step 1: The first data processing step loads a Parquet file to create a DataFrame with a "Telephone number" column.

Step 2: Data stored in the "Telephone" column is cleaned and transformed into three columns to separate the country code, the area code, and the local phone number.

Step 3: A data processing step creates a second DataFrame with other information, such as age, from a database.

Step 4: These two DataFrames are joined and loaded into the data warehouse for further analysis.

- ◉ Step 1, Step 2, Step 3, Step 4
- ○ Step 1, Step 4, Step 3, Step 2
- ○ Step 4, Step 2, Step 1, Step 3
- ○ Step 1, Step 3, Step 2, Step 4

7. Select the answers that define and describe Graph Theory. (Multiple answers) **1 point**

- ☐ Graph theory for Apache Spark is the study of graphs generated from parametric specifications.
- ☐ The graph is a construct that contains an X, Y, and Z-axis.
- ☑ The graph is a construct that contains a set of vertices with pairwise edges that connect one vertex to another.
- ☑ Graph theory is the mathematical study of modeling pairwise relationships between objects.

8. Select the options that define watermarking. (Multiple answers) **1 point**

- ☑ Enables the inclusion of late-arriving data stream processing
- ☑ Is the process that manages late data
- ☐ Is the process that manages and tags first-arriving data
- ☑ Updates results after initial data processing.

9. Select the statements that are true about using GraphFrames. (Multiple Answers) **1 point**

- ☑ Provides one DataFrame for graph vertices and one DataFrame for edges that can be used with SparkSQL for analysis
- ☑ Performs Motif finding, which searches the graph for structural patterns. Motif finding is supported in GraphFrames with the `find()` method that uses domain specific language (DSL) to specify the search query in terms of edges and vertices.
- ☐ Is ideal for modeling data with connecting relationships and computes relationship strength and direction
- ☑ Comes with popular built-in graph algorithms for use with the edge and vertex DataFrames

10. Select the built-in data sources from which Spark can extract data. **1 point**

- ☐ Microsoft Excel
- ☑ JDBC
- ☑ Apache ORC
- ☑ Parquet

Upgrade to submit