Welcome.

We'll start at the top of the hour.

In the meantime, consider your experience with and knowledge of the following:

1. **Apache Spark Architecture**
2. **Apache Spark DataFrames**

DATA+AI SUMMIT 2021

FORMERLY SPARK+AI SUMMIT

Certification Prep

Databricks Certified Associate Developer for Apache Spark

Instructor - Mark Roepke

Course objectives

- Understand the learning context behind the Databricks Certified Associate Developer for Apache Spark exam (the exam).
- Describe the topics covered in the exam.
- Describe the format and structure of the exam.
- Apply practical test-taking strategies to answer example questions similar to those of the exam.
- Highlight resources that can be used to learn the material covered in the exam.

What this course will not do

- **Teach the actual content assessed by the exam.**
- Provide answers to exam questions.

Agenda

1. Certification Philosophy
2. Exam Topics
3. Exam Format and Structure
4. Exam Questions
5. Exam Study Resources

Certification Philosophy

Section Objective

Understand the learning context behind the Databricks Certified Associate Developer for Apache Spark exam (the exam).

Databricks Certification Program

Certification program history

2014

First Apache Spark certification exam

Joint venture with O'Reilly to launch the first Apache Spark certification exam that focused on Apache Spark 1.6.

2018

Certification exam for Apache Spark 2

Databricks Academy released a new certification for Apache Spark 2 that gave students the choice between using Scala and Python.

2019

Databricks Certified Associate Developer for Apache Spark 2.4

The Databricks Certified Associate Developer for Apache Spark exam launched with a focus on the general aspects of Spark.

2020

Databricks Certified program growth

Databricks Academy scaled its certification program to include exams for different data-focused roles, launching multiple certifications for Apache Spark 2.4 and 3.0.

Types of assessments

Accreditations

- Public-facing
- Course-aligned
- Unproctored
- Lower stakes
- Multiple-choice

Partner Badges

- Partner-facing
- Demonstrate capabilities to prospective clients
- Performance-based

Certifications

- Public-facing
- Role-aligned
- Tiered
- Proctored
- Higher-stakes
- Multiple-choice

Available certifications

Associate Developer for Apache Spark

- This course!
- Apache Spark
- Assesses understanding of Spark Architecture and the ability to use Spark DataFrame API

Azure Databricks Associate Platform Administrator

- Azure-specific
- Databricks platform
- Assesses understanding of basics in network infrastructure and security, identity and access, clusters, and automation.

Associate ML Practitioner for Apache Spark

- Apache Spark
- Assesses understanding of and ability to apply machine learning techniques using the Spark MLlib library

Professional Data Scientist

- Tool-agnostic
- Professional-level
- Assesses the understanding of the basics of ML, steps in the ML lifecycle, understanding of ML algorithms, and basics of MLflow.

Future plans

In the future, Databricks Academy is considering the release of additional certifications:

- Professional Data Engineer
- AWS/Google Cloud Associate Platform Administrator
- Professional Machine Learning Engineer

Databricks Certified Associate Developer for Apache Spark Overview

Data roles

- **SQL Analyst** - Explore and analyze data to answer organizational questions using SQL and visualization
- **Data Scientist** - Build machine learning models, optimization solutions, etc., to answer complex organizational questions
- **Machine Learning Engineer** - Deploy, monitor, and maintain already-built machine learning models
- **Data Engineer** - Build data pipelines to move and clean organizational data
- **Data Architect** - Design data systems by selecting the optimal technologies and settings
- **Platform Administrator** - Maintain the Databricks platform for an organization

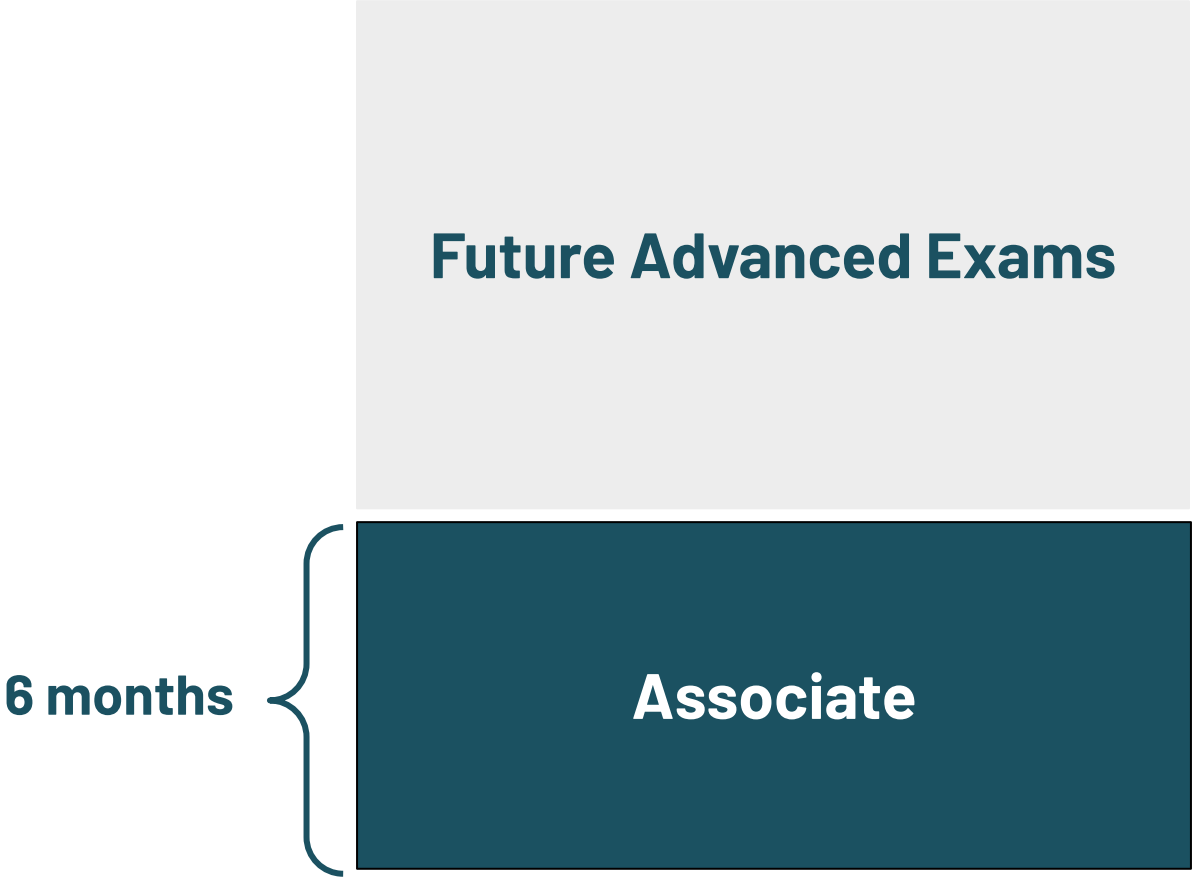
Target audience

The Databricks Certified Associate Developer for Apache Spark exam assesses the beginner Apache Spark skills needed by the following data roles:

- Data Scientist
- Machine Learning Engineer
- Data Engineer
- Data Architect

Definition of “Associate”

- Entry-level certification
- Assess candidates at a level equivalent to **six months of experience** with the certification’s topic



Associate Developer Expectations

Therefore, the following is expected of an Associate-level developer:

- Understanding of the basics of the Apache Spark architecture
- Ability to perform basic data manipulations using the Apache Spark DataFrame API
- Ability to read and write non-streaming data using Apache Spark
- Ability to apply basic scaling and debugging mechanisms for Apache Spark clusters

Out-of-scope

And the following is **not** expected of an Associate-level developer:

- Ability to tune Apache Spark jobs
- Memorization of the Apache Spark APIs
- Ability to create data visualizations
- Ability to build, evaluate, deploy, and manage machine learning models
- Understanding of data engineering and machine learning pipelines
- Ability to set up real-time data streams

Knowledge Check

There is a required prerequisite certification exam prior to the Databricks Certified Associate Developer for Apache Spark exam.

- True
- False

Select the roles aligned to the Databricks Certified Associate Developer for Apache Spark exam.

- SQL Analyst
- Data Scientist
- Machine Learning Engineer
- Data Engineer
- Data Architect
- Platform Administrator

Select the expectations driving the design of the Databricks Certified Associate Developer for Apache Spark exam.

- Ability to tune Apache Spark jobs
- Memorization of the Apache Spark APIs
- Understanding of the basics of the Apache Spark architecture
- Ability to perform basic data manipulations using the Apache Spark DataFrame API
- Ability to create data visualizations
- Ability to read and write non-streaming data using Apache Spark
- Ability to apply basic scaling and debugging mechanisms for Apache Spark clusters
- Ability to set up real-time data streams

Exam Topics

Section Objective

Describe the topics covered in the exam.

High-level Topics

High-level Exam Topics

Apache Spark Architecture

- Cluster-computing framework
- Describes how data is partitioned, processed, etc.

Apache Spark DataFrame API

- Fundamental user-facing data structure of Apache Spark
- Used to manipulate data using common data manipulation terminology

Spark Architecture Basics

Spark Architecture Concepts

- Cluster architecture: nodes, drivers, workers, executors, slots
- Spark execution hierarchy: applications, jobs, stages, tasks
- Shuffling
- Partitioning
- Lazy evaluation
- Transformations vs. actions
- Narrow vs. wide transformations

Sparks Architecture Applications

- Execution deployment modes: cluster, client, local
- Stability and fault tolerance
- Garbage collection
- Out-of-memory errors
- Storage levels
- Repartitioning
- Coalescing
- Broadcasting
- DataFrames

Spark DataFrame API Basics

Spark DataFrame API Applications

- Subsetting DataFrames
- Column manipulation
- String manipulation
- Performance-based operations
- Combining DataFrames
- Reading/writing DataFrames
- Working with dates
- Aggregations
- Sorting
- Missing values
- Typed UDFs
- Value extraction
- Sampling
- Working with rows

Minimally-qualified Candidate

The minimally-qualified candidate should:

- Have a basic understanding of the Spark architecture
- Be able to apply the Spark DataFrame API to complete individual data manipulation tasks:
  - Selecting, renaming and manipulating columns
  - Filtering, dropping, sorting, and aggregating rows
  - Joining, reading, writing and partitioning DataFrames
  - Working with UDFs and Spark SQL functions

Self-assessment on Topics

Self-assessment Activity

On the next two slides, there will be a series of statements **describing an objective/task relating to the topics covered by the exam**.

For each statement, select **one** of the following:

- Very underprepared
- Somewhat underprepared
- Prepared

based on how **your ability to complete that objective/task**.

Self-assessment on Spark Architecture

- Describe the difference between the Spark driver and a Spark executor.
- Form a hierarchy of Spark jobs, tasks, stages, and applications.
- Describe what causes data to shuffle.
- Describe the difference between transformations and actions.
- Describe the difference between local and cluster execution modes.
- Determine what happens to the Spark application if the driver shuts down.
- Describe garbage collection strategies in Spark.
- Describe the advantages and disadvantages of caching data at various storage levels.
- Describe how DataFrames are repartitioned.

Self-assessment on Spark DataFrame API

- Select a subset of columns from a DataFrame.
- Filter a subset of rows from a DataFrame based on two logical filtering criteria.
- Cast a column from a numeric type to a string type.
- Create a new DataFrame column by mathematically combining two existing columns.
- Split a string DataFrame column into two columns based on a regular expression.
- Cache a DataFrame to a specific storage level.
- Aggregate data to find the mean of a column by group.
- Write a DataFrame to disk.
- Extract the month from a DataFrame column of date type.
- Create a UDF to use in a Spark SQL statement.

Exam Format and Structure

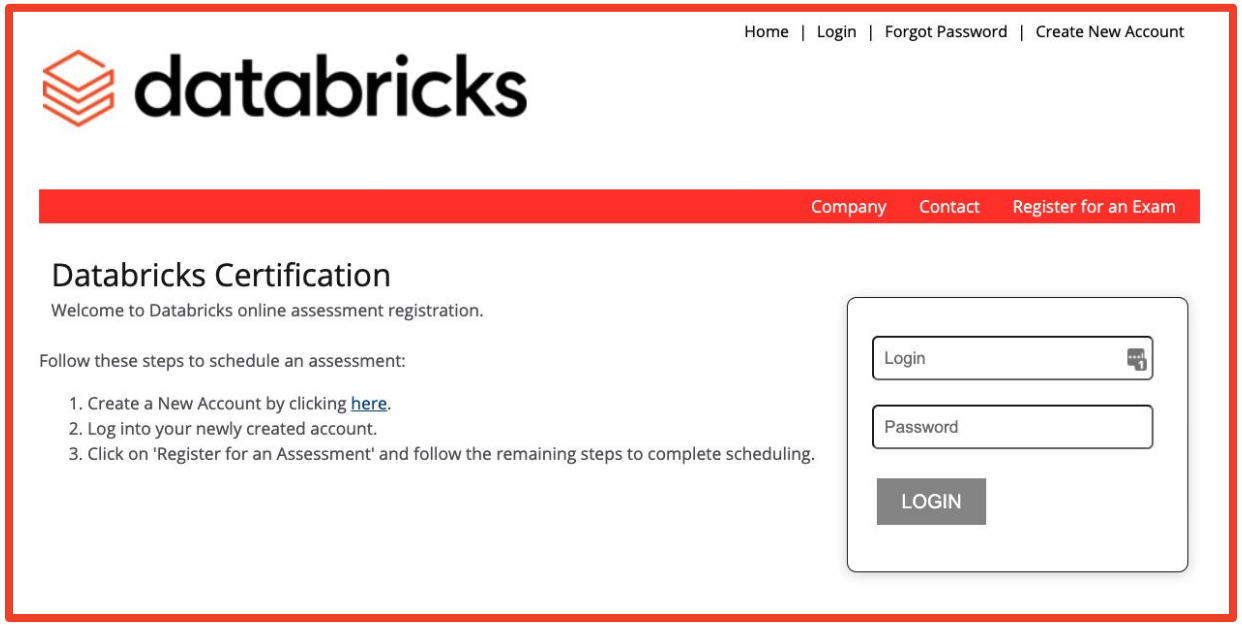
Section Objective

Describe the format and structure of the exam.

Exam Delivery

Kryterion

- Databricks Academy certifications are offered through **Kryterion's Webassessor platform**.
- Webassessor is a simple, scalable assessment solution resulting in an easy test-taking experience.



Proctoring Details

- During the exam, you will be:
  - **Monitored via webcam** by a Webasssessor proctor.
  - Asked to provide **valid, photo-based identification.**
- The proctor will:
  - Monitor you during the exam.
  - Answer any exam delivery questions you might have.
  - Provide technical support.
- The proctor will **not** provide assistance on the content of the exam.

Exam Details

Basic Exam Details

- Exam offered in **Python or Scala**
- Exam fee = **$200**
- Exam *retake* fee = **$200**
- Time allotted to complete exam = **2 hours**
- Passing scores = At least **70%**
- Exam retake fee = **$200**
- More info. on the Certification FAQ page: https://academy.databricks.com/training-faq#cert-faq.

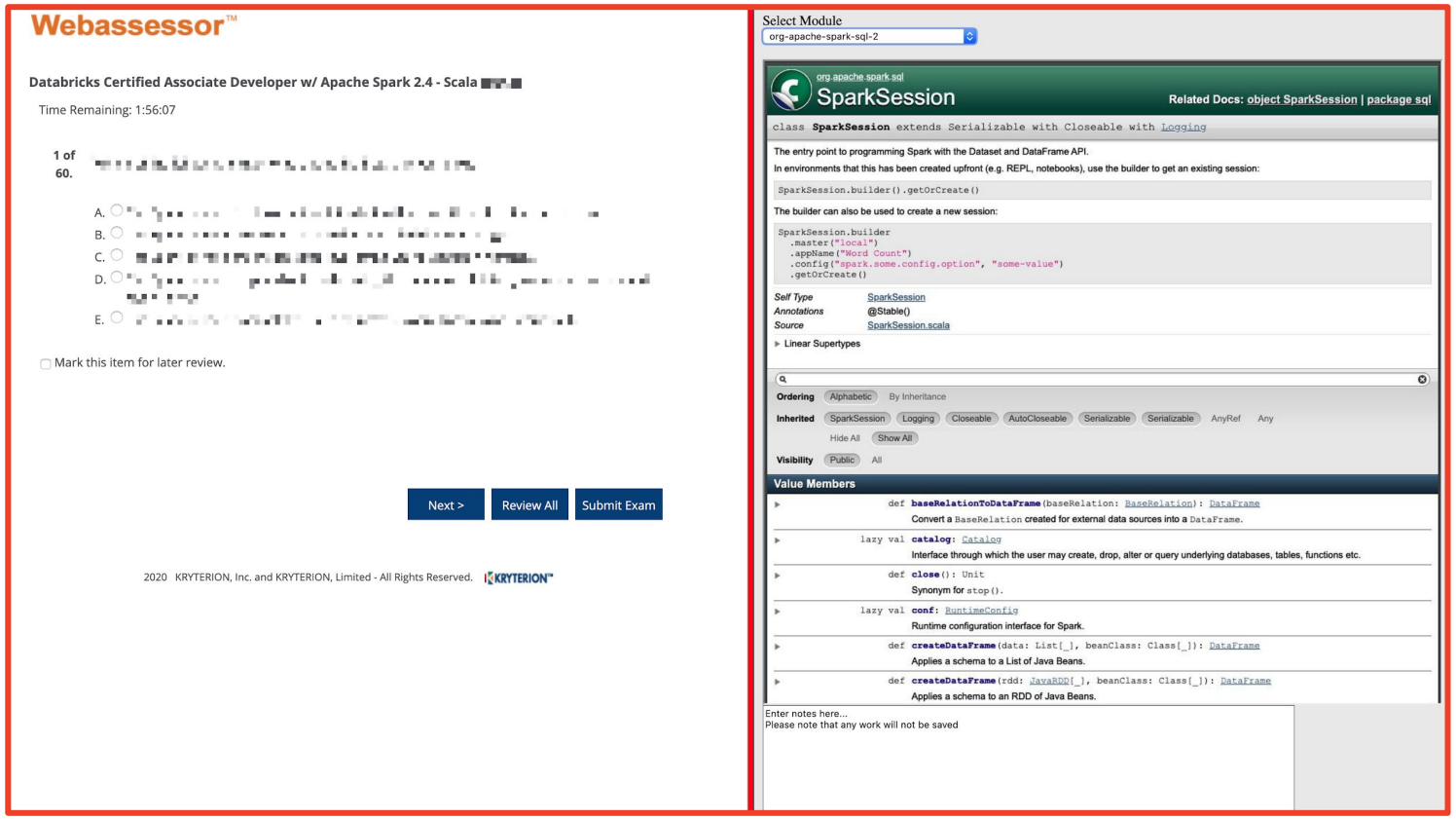
Exam Question and Topic Distribution

- Exam questions are distributed into three categories:
  - Spark Architecture: Conceptual understanding (**~17%**)
  - Spark Architecture: Applied understanding (**~11%**)
  - Spark DataFrame API Applications (**~72%**)

Test Aids

Test Aids

- Spark docs (PDF)
- Notepad



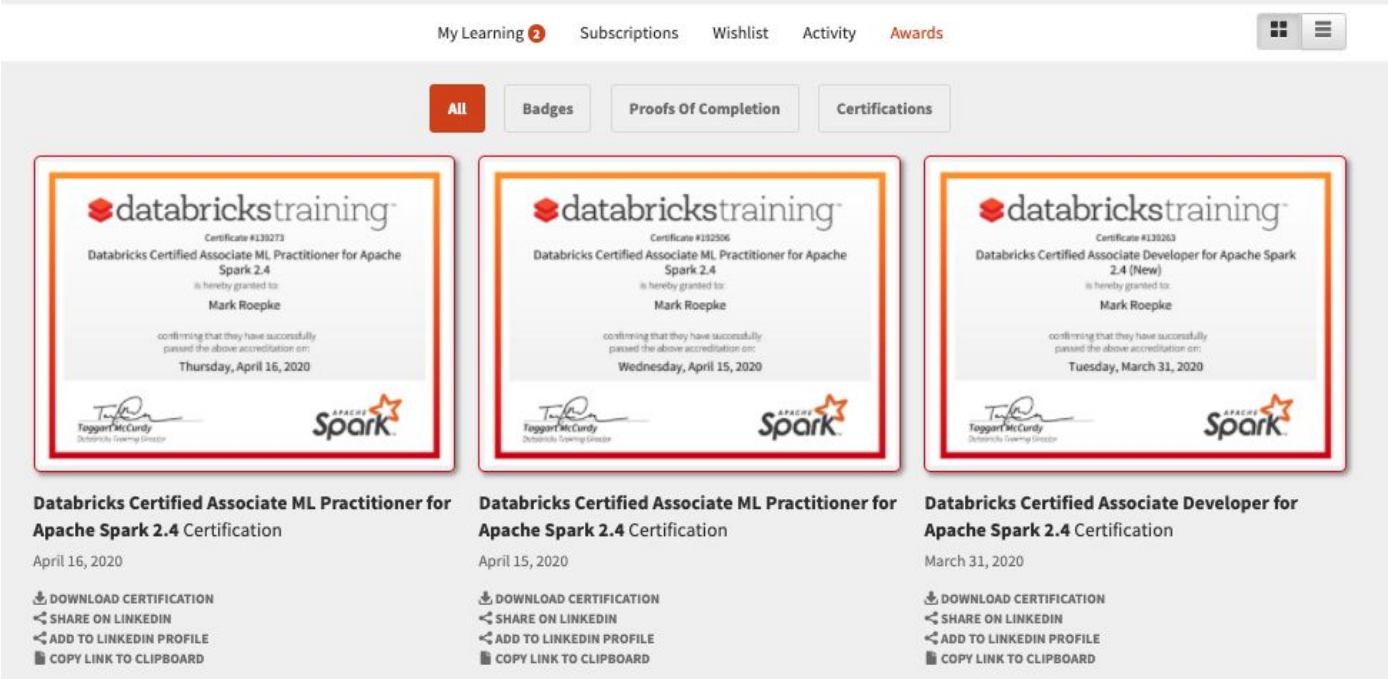
Grading and Certification

Exam Grading

- Certification exams are automatically graded.
- Following the exam, the proctor's session notes and the recorded grade will be reviewed by Databricks Academy,
- It will take about **one week** for you to find out whether or not you passed the exam.

Certificate Awarding Process

- If it's been determined that you've passed the exam, **your certificate will be awarded via Databricks Academy.**



Knowledge Check

All attempts of the Databricks Certified Associate Developer for Apache Spark exam will be proctored virtually.

- True
- False

Which of the following is the price of the Databricks Certified Associate Developer for Apache Spark certification exam?

A. $200
B. $300
C. $300, but with free retakes
D. $200, but with free retakes

The passing threshold for the exam is 70 percent.

- True
- False

Students can retake the exam as many times as they would like.

- True
- False

Select all of the following resources that will be available during the exam.

A. The Spark API documentation
B. A digital notepad
C. Paper and pencil
D. A single, index card of pre-written notes
E. A running Spark session

I will receive my results within one week of completing the exam.

- True
- False

Exam Questions

Section Objective

Apply practical test-taking strategies to answer example questions similar to those of the exam.

Types of Questions

Exam Question and Topic Distribution

- Exam questions are distributed into three categories:
  - Spark Architecture: Conceptual understanding (~17%)
  - Spark Architecture: Applied understanding (~11%)
  - Spark DataFrame API Applications (~72%)

- All of the questions in all categories are **multiple-choice questions** – this means there's only one correct answer for each question.

Spark Architecture Question Types

- **Definitions:** what something is or does
- **Relationships:** how something compares to or is related to something else
- **Results:** If _____ occurs, ...
- **Classification:** in which category does something belong
- **Cluster Configurations:** based on this cluster configuration, ...

Example Definition Question

Which of the following describes a worker node?

a. Worker nodes are the nodes of a cluster that perform computations.
b. Worker nodes are synonymous with executors.
c. Worker nodes always have a one-to-one relationship with executors.
d. Worker nodes are the most granular level of execution in the Spark execution hierarchy.
e. Worker nodes are the most coarse level of execution in the Spark execution hierarchy.

Example Relationship Question

Which of the following describes the relationship between worker nodes and executors?

a. An executor is a Java Virtual Machine (JVM) running on a worker node.
b. A worker node is a Java Virtual Machine (JVM) running on an executor.
c. There are always more worker nodes than executors.
d. There are always the same number of executors and worker nodes.
e. Executors and worker nodes are not related.

Example Results Question

If Spark is running in cluster mode, which of the following statements about nodes is incorrect?

a. There is a single worker node that contains the Spark driver and the executors.
b. The Spark driver runs in its own non-worker node without any executors.
c. Each executor is a running JVM inside of a worker node.
d. There is always more than one node.
e. There might be more executors than total nodes or more total nodes than executors.

Example Classification Question

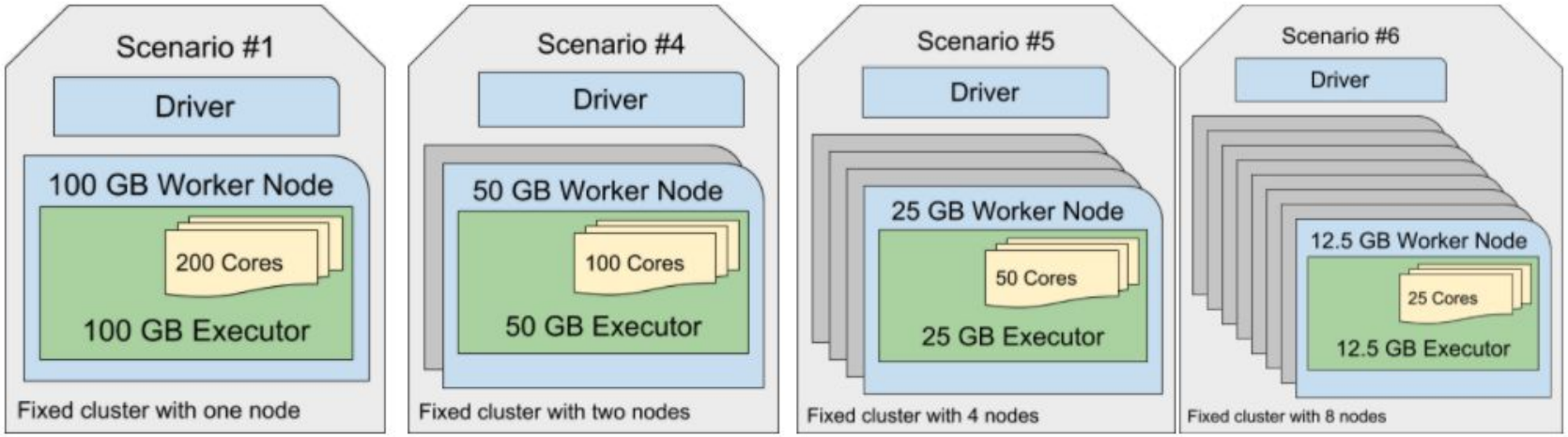
Which of the following DataFrame operations is always classified as a narrow transformation?

a. `DataFrame.select()`
b. `DataFrame.sort()`
c. `DataFrame.distinct()`
d. `DataFrame.join()`
e. `DataFrame.repartition()`

Example Cluster Configuration Question

Which of the following cluster configurations is most likely to result in the greatest number of shuffles?

a. Scenario #1
b. Scenario #4
c. Scenario #5
d. Scenario #6
e. More information is needed to determine an answer.



Note: each configuration has roughly the same compute power using 100GB of RAM and 200 cores.

DataFrame API Question Types

- For the questions assessing Spark's DataFrame API, there are a few formats:
  - **Operation identification** – which operation does _____
  - **Code block comparison** – which of the following code blocks correctly _____
  - **Error identification** – identify the error in the code block
  - **Fill-in-the-blank** – complete the code block by filling in the blanks
  - **Ordering lines of code** – place the lines of code in order to correctly _____

Example Operation Identification Question

Which of the following operations can be used to create a new DataFrame with a new column and all previously existing columns form an existing DataFrame?

a. `DataFrame.withColumn()`
b. `DataFrame.drop()`
c. `DataFrame.withColumnRenamed()`
d. `DataFrame.head()`
e. `DataFrame.filter()`

Example Code Block Comparison Question

Which of the following code blocks returns a DataFrame with a new column `aSquared` and all previously existing columns from DataFrame `df`?

a. `df.withColumn("aSquared", col("a") * col("a"))`
b. `df.withColumnRenamed("aSquared", col("a") * col("a"))`
c. `df.select("aSquared")`
d. `df.withColumn(col("a") * col("a"), "aSquared")`
e. `df.withColumnRenamed(col("a") * col("a"), "aSquared")`

Example Error Identification Question

The code block shown below contains an error. The code block is intended to return a DataFrame with a new column `aSquared` and all previously existing columns from DataFrame `df`. Identify the error.

Code block: `df.withColumn(col("a") * col("a"), "aSquared")`

a. The arguments to `df.withColumn()` are provided in reverse order. "`aSquared`" should be first, and `col("a") * col("a")` should be second.

b. The `df.withColumn()` operation does not create new columns. The `df.newColumn()` operation should be used instead.

c. The argument "`aSquared`" must be wrapped in the `col()` function because it is a column name.

d. The `withColumn()` operation is not a DataFrame method. It should be called on its own with the first argument being `df`.

e. The `df.withColumn()` operation does not create new columns. The `df.withColumnRenamed()` operation should be used instead.

Example Fill-in-the-blank Question

The code block shown below should return a DataFrame with a new column `aSquared` and all previously existing columns from DataFrame `df`. Choose the response that correctly fills in the numbered blanks within the code block to complete this task.

Code block: `df.__1__(__2__, __3__)`

A
1. withColumn
2. "aSquared"
3. col("a") * col("a")

B
1. withColumnRenamed
2. "aSquared"
3. col("a") * col("a")

C
1. withColumn
2. col("aSquared")
3. col("a") * col("a")

D
1. withColumn
2. "aSquared"
3. "a" * "a"

E
1. withColumnRenamed
2. "aSquared"
3. "a" * "a"

Example Order-lines-of-code Question

In what order should the below lines of code be run in order to return a DataFrame with a new column `aSquared` and all previously existing columns from DataFrame `df`?

```
1. df
2. .withColumn("aSquared", "a" * "a")
3. .withColumn("aSquared", col("a") * col("a"))
4. DataFrame
5. .withColumn(col("aSquared"), col("a") * col("a")
```

a. 1, 3

b. 1, 2

c. 1, 5

d. 4, 2

e. 4, 3

Test-taking Strategies

Preparation Strategies

- **Be prepared** - There is no substitute for knowing the material. Use the self-assessment to identify potential knowledge gaps and close those gaps.
- **Give yourself time** - Schedule the exam far enough in advance to give yourself time to prepare.
- **Familiarize yourself with Spark documentation** - The exam requires use of the Spark documentation. Become familiar with navigating the documentation and using it to answer specific questions.
- **Practice debugging code** - Look at existing code blocks that are resulting in errors. Identify where the error is located, why it's causing an error, and what can be done to resolve the issue.

Multiple Choice Strategies

- **Read carefully** - The questions are detailed. Read them, and all of the responses, thoroughly.
- **Answer the question mentally** - Prior to reviewing the responses, answer the question mentally to familiarize yourself with what to expect as a correct answer.
- **Eliminate incorrect responses** - If you're sure that a response is incorrect, eliminate it to narrow down your potential answers.
- **Keep moving** - All of the questions are worth the same number of points. Do not get stuck on one question. Mark any troubling questions for review and revisit them at the end.

Code-based Multiple Choice Strategies

- **Read carefully** - Some of the questions ask for very specific tasks. Be sure that you understand the task, and review the Spark API documentation to ensure the correct operations and arguments are being used.
- **Visualize the result** - Visualize what the result of the code should look like. Is there a new DataFrame? How is it changed?
- **Focus on the logic** - While there are small differences in the code-based responses, there are not intended to be typos. Any differences between code-based responses are logical.
- **Write the code** - Don't be afraid to write code in the digital notepad test aid. That's what it's there for, and writing code is more natural.
- **Mentally run the code** - While there is no Spark session in the exam, mentally run any code. Will it cause an error? Will it do what it's supposed to do?

Applying Test-taking Strategies

Applying Test-taking Strategies to Spark Architecture Questions

Which of the following describes a worker node?

A. Worker nodes are the nodes of a cluster that perform computations.
B. Worker nodes are synonymous with executors.
C. Worker nodes always have a one-to-one relationship with executors.
D. Worker nodes are the most granular level of execution in the Spark execution hierarchy.
E. Worker nodes are the most coarse level of execution in the Spark execution hierarchy.

Applying Test-taking Strategies to Spark DataFrame API Questions

The code block shown below contains an error. The code block is intended to return a DataFrame with a new column `aSquared` and all previously existing columns from DataFrame `df`. Identify the error.

Code block: df.withColumn(col("a") * col("a"), "aSquared")

A. The arguments to df.withColumn() are provided in reverse order. "aSquared" should be first, and col("a") * col("a") should be second.

B. The df.withColumn() operation does not create new columns. The df.newColumn() operation should be used instead.

C. The argument "aSquared" must be wrapped in the col() function because it is a column name.

D. The withColumn() operation is not a DataFrame method. It should be called on its own with the first argument being df.

E. The df.withColumn() operation does not create new columns. The df.withColumnRenamed() operation should be used instead.

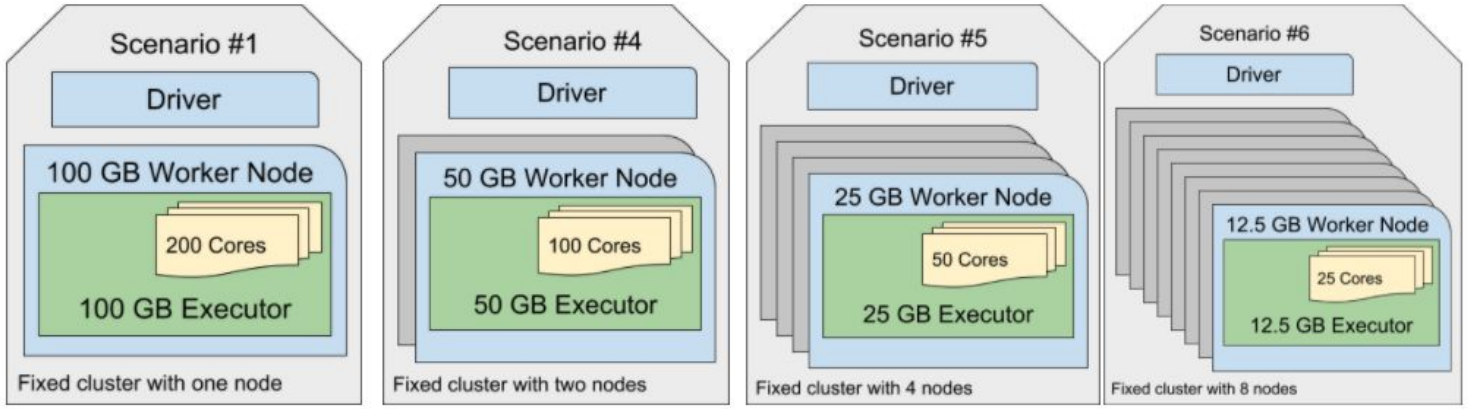
Strategy Identification Exercise

Which of the following code blocks returns a DataFrame with a new column aSquared and all previously existing columns from DataFrame df?

A. Worker nodes are the nodes of a cluster that perform computations.
B. Worker nodes are synonymous with executors.
C. Worker nodes always have a one-to-one relationship with executors.
D. Worker nodes are the most granular level of execution in the Spark execution hierarchy.
E. Worker nodes are the most coarse level of execution in the Spark execution hierarchy.

Which of the following cluster configurations is mostly likely to result in the greatest number of shuffles?

A. Scenario 1
B. Scenario 4
C. Scenario 5
D. Scenario 6
E. More information is needed to determine an answer.



Note: each configuration has roughly the same compute power using 100GB of RAM and 200 cores.

Exam Study Resources

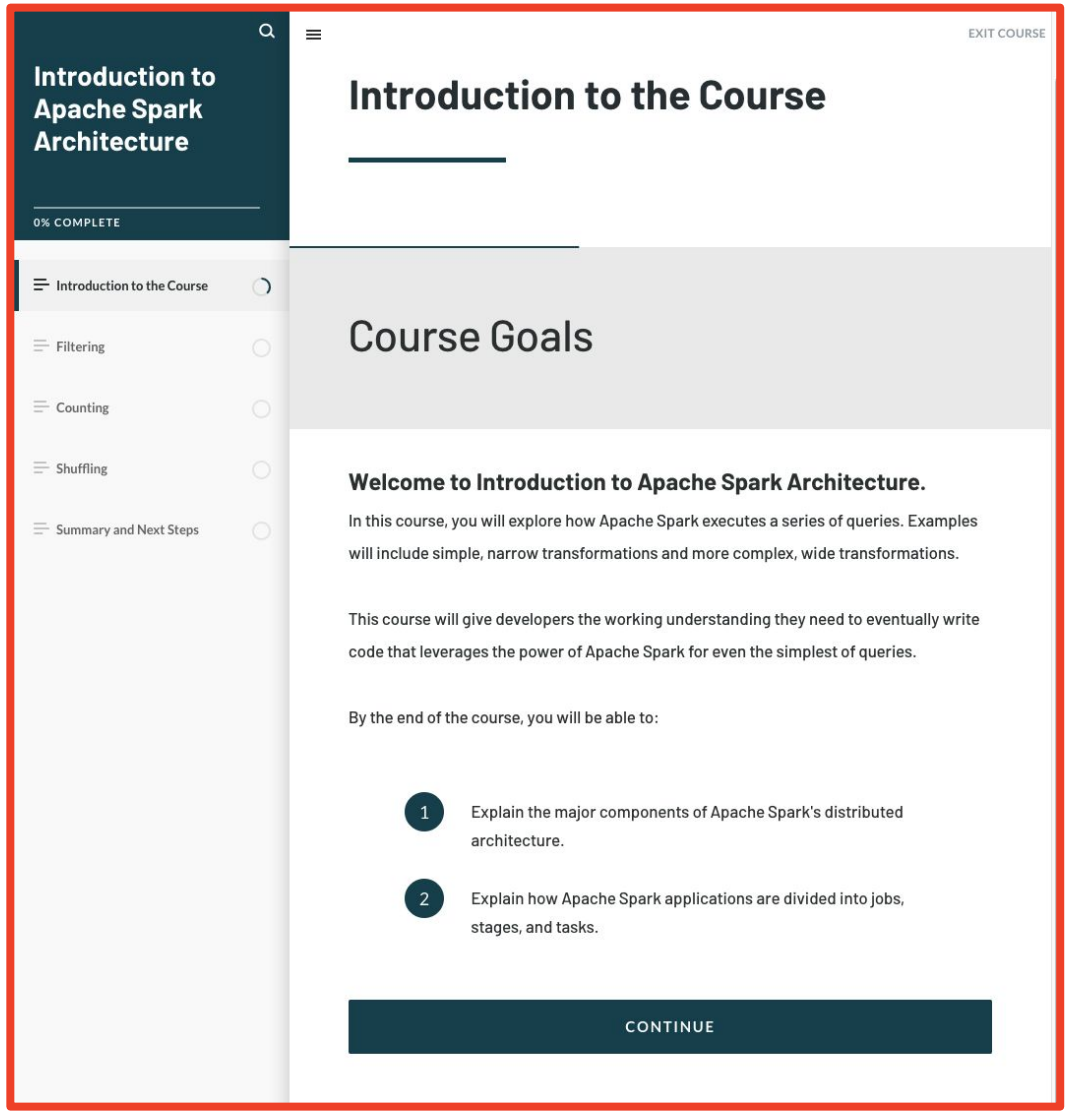
Section Objective

Highlight resources that can be used to learn the material covered in the exam.

Databricks Academy

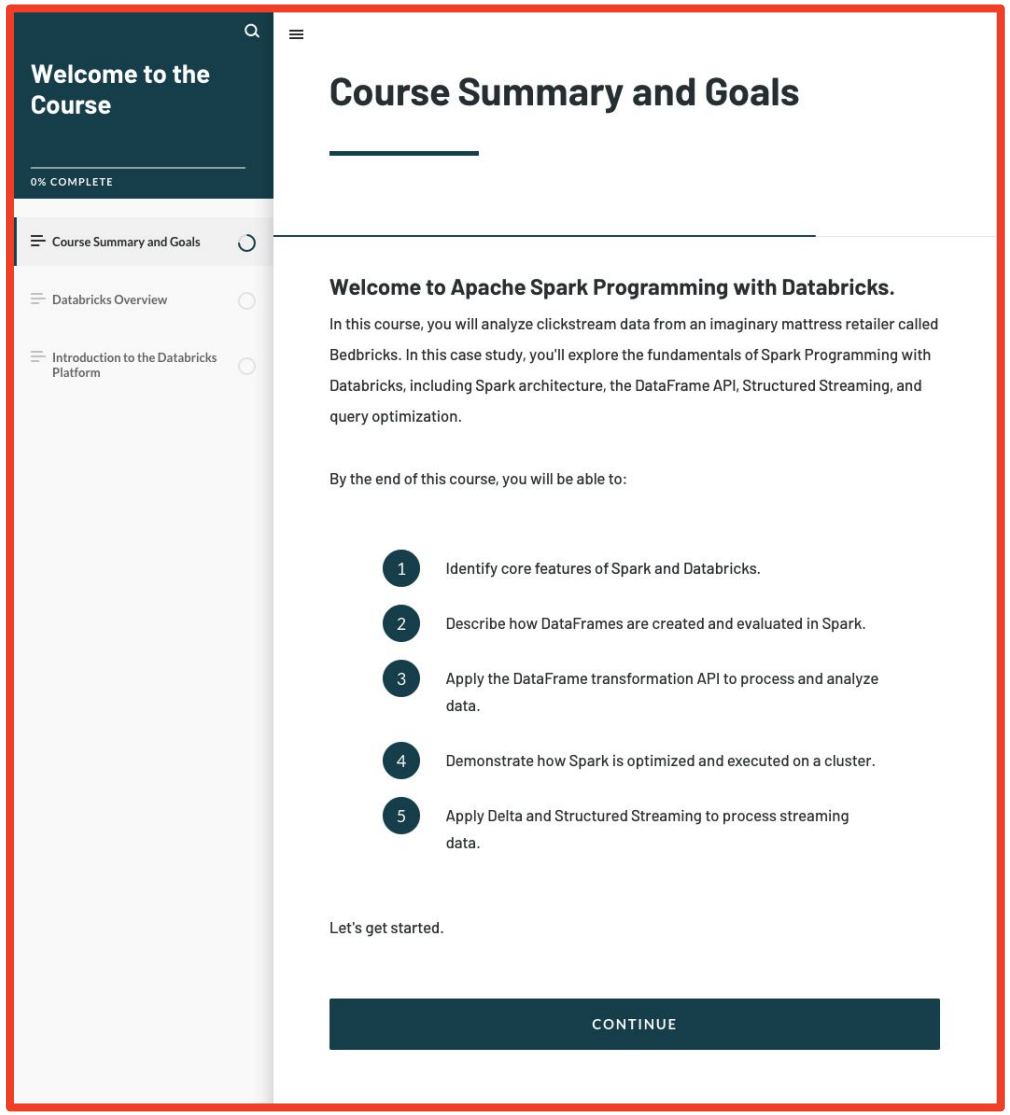
Introduction to Apache Spark Architecture

- Free for customers/partners
- Self-paced
- Will prepare you for the **Spark architecture** portion of the exam
- More info: https://academy.databricks.com/elearning/INT-SPARCH-v1-SP



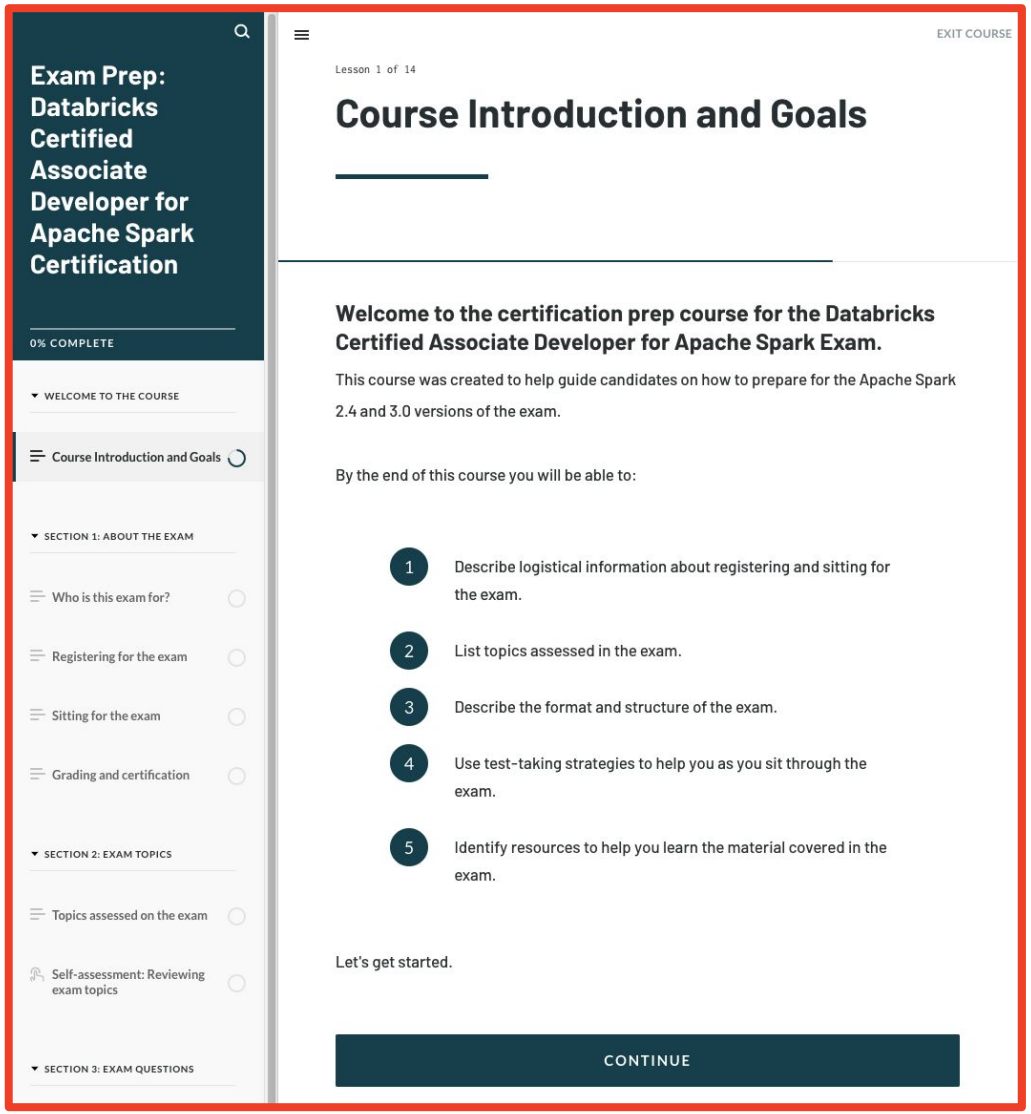
Apache Spark Programming with Databricks

- Self-paced
  - Free for customers/partners
- Public and private instructor-led
- Will prepare you for the **Spark DataFrame API** portion of the exam
- More info: https://academy.databricks.com/instructor-led-training/apache-spark-programming



Certification Prep Course

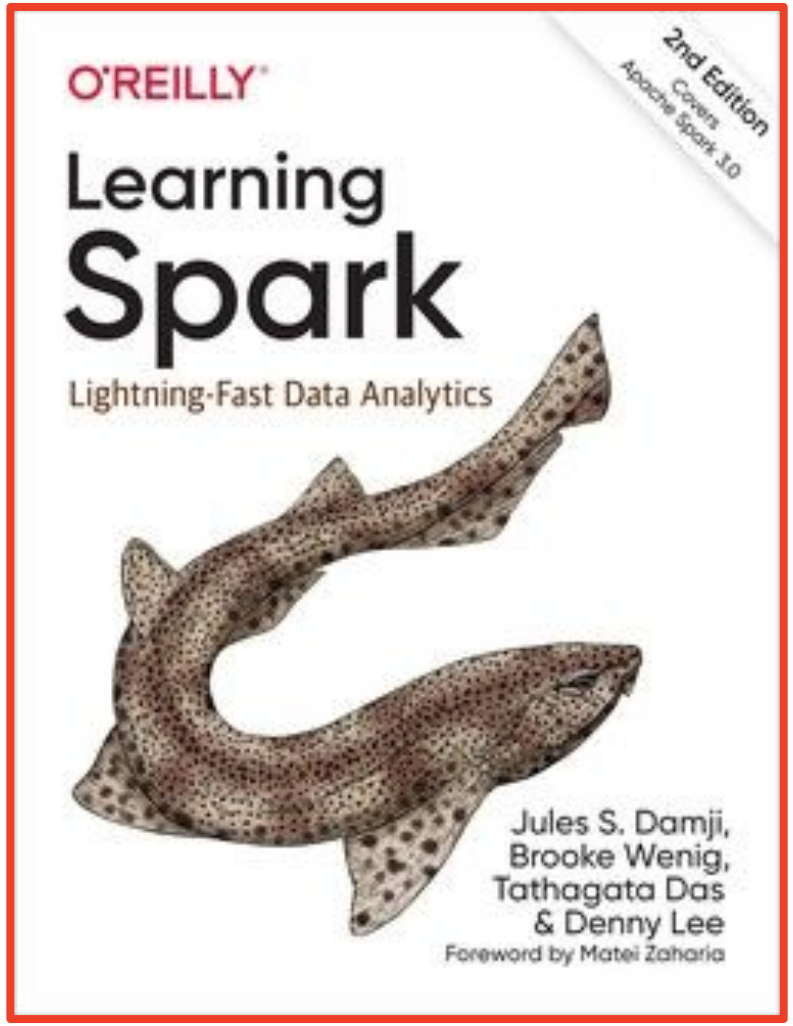
- Self-paced
  - Free for customers/partners
- Data+AI Summit (this course!)
- Will provide essential info about the exam
- More info: https://academy.databricks.com/elearning/INT-DCAD-v2-SP



Textbooks

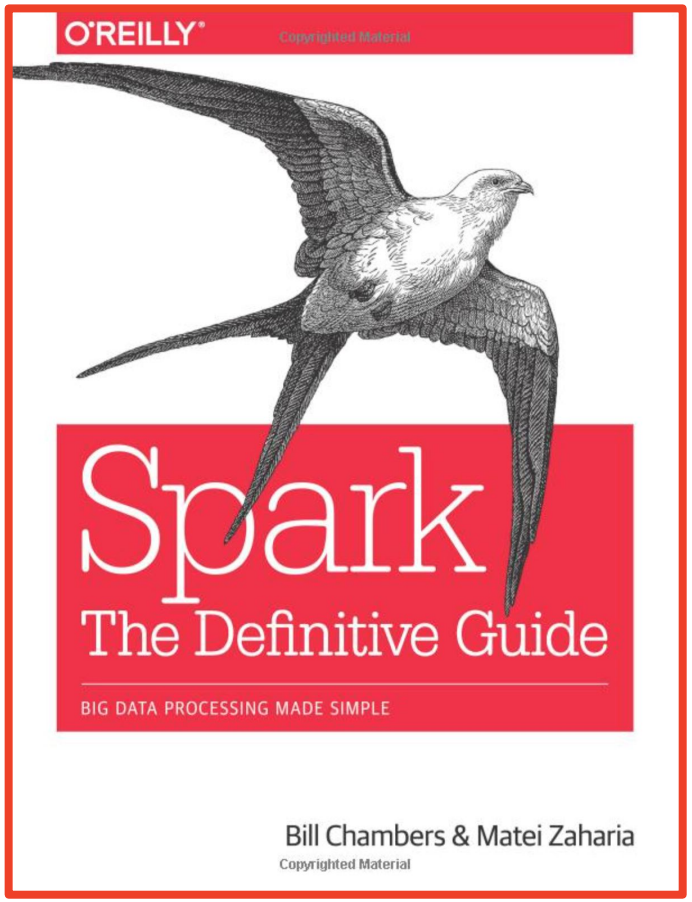
Learning Spark, 2nd Edition

- Comprehensive overview of exam material
- Sections I, III, IV, V, XII should be particularly helpful
- Updated for Spark 3.0 (XII)
- Link: https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/



Spark: The Definitive Guide

- Comprehensive overview of exam material
- Sections I, II, IV should be particularly helpful
- Does not cover Spark 3.0 specifics
- Link: https://www.oreilly.com/library/view/spark-the-definitive/9781491912201/.



Practice Exam

PRACTICE EXAM!

- Full, 60-question practice exam.
- Made up of retired questions.
- Representative of the actual exam.
- Answers provided at the end.
- Access here: files.training.databricks.com/assessments/practice-exams/PracticeExam-DCAD3.pdf

databricks Academy

Practice Exam

Databricks Certified Associate Developer for Apache Spark 3.0

Overview

This is a practice exam for the Databricks Certified Associate Developer for Apache Spark 3.0 exam. The questions here are retired questions from the actual exam that are representative of the questions one will receive while taking the actual exam. After taking this practice exam, one should know what to expect while taking the actual Associate Developer for Apache Spark 3.0 exam.

Just like the actual exam, it contains 60 multiple-choice questions. Each of these questions has one correct answer. The correct answer for each question is listed at the bottom in the **Correct Answers** section.

There are a few more things to be aware of:

1. This practice exam is for the Python version of the actual exam, but it's incredibly similar to the Scala version of the actual exam, as well. There is no practice exam for the Scala version of the actual exam due to the similarity between the two.
2. There is a two-hour time limit to take the actual exam.
3. In order to pass the actual exam, testers will need to correctly answer at least 42 of the 60 questions.
4. During the actual exam, testers will be able to reference the Apache Spark documentation. Please use the documentation while taking this practice exam.
5. During the actual exam, testers will not be able to test code in a Spark session. Please do not use a Spark session when taking this practice exam.
6. These questions are representative of questions that are on the actual exam, but they are no longer on the actual exam.

We’ve reached the end of our course…

At this point, you should be able to:

- Understand the learning context behind the Databricks Certified Associate Developer for Apache Spark exam (the exam).
- Describe the topics covered in the exam.
- Describe the format and structure of the exam.
- Apply practical test-taking strategies to answer example questions similar to those of the exam.
- Highlight resources that can be used to learn the material covered in the exam.

Next steps

1. Review the results of your self-assessment on the topics covered by the exam.
2. Better learn topics for which you're underprepared using the provided resources.
3. Take the practice exam and calculate your score.
4. Register for and take the exam!

Feedback

Your feedback is important to us.

Don’t forget to rate and review the sessions.

DATA+AI
SUMMIT 2021
FORMERLY SPARK+AI SUMMIT

Thank you!