SPARK+AI
SUMMIT EUROPE

Databricks Developer Certification: for Apache Spark 2.x

with Lyuben Todorov & Jamie Murray

Wi-Fi SSID: Spark+AISummit
Password: bigdata2018

No Hot Spots Please!

Introductions

Lyuben Todorov (lyubent123@gmail.com)

- Databricks Certified Instructor
- Distributed Systems Engineer

Jamie Murray (jamie@databricks.com)

- Business Operation Coordinator
- Coordinator of the Databricks Certified Developer exam

Class Schedule

Training	9:00 AM to 10:30 AM
Break	10:30 AM to 10:50 AM
Training	10:50 AM to 12:00 PM
Lunch	12:00 PM to 1:00 PM
Meet Up	6:00 PM to 8:30 PM

Exam Schedule

Tuesday	**12:15 PM to 3:00 PM**
Wednesday	**10:30 AM to 3:00 PM (After keynote)**
Thursday	**10:30 AM to 3:00 PM**
	Last sign-up at 3 PM Thursday.
Online	http://databricks.link/spark-certification

History

Co-branded Spark Certification

- September 2014 O’Reilly and Databricks announced a joint program
- Consisted of questions covering multiple languages
- Became out of date with updates and new product releases

Databricks Certified Developer - Apache™ Spark 2.x

- Introduced in January 2018
- Produced independently by Databricks Academy
- Exam bifurcated to accommodate primary developer knowledge and experience
- Available in Scala and Python
- All questions aligned to latest release

Today’s Agenda

1. @Summit Requirements
2. Test Voucher Usage
3. Post-Summit Requirements
4. Key Points About the Exam
5. Certification Exam Scope
6. Format of Exam Questions
7. Exam Topic Areas:
   - Spark Architecture and Run-time Behavior
   - Spark SQL and DataFrame/DataSet Manipulation
   - RDDs and Low-Level APIs
   - Structured Streaming
   - Machine Learning
   - GraphFrames
   - Key API Classes
8. Spark Study Resources
9. Questions & Answers
10. Creating An Account
11. Purchase An Exam
12. Taking The Practice Exam

@Summit Requirements (1/2)

What you will need

1. You will need to create an account on the Databricks Webassessor Portal
   - We will do this together after the lecture

2. A test voucher code
   - To be handed out during this class
   - You will need this to select & purchase the exam

3. The authorization code
   - Provided after the exam is purchased
   - The proctor will need this code to launch the exam

@Summit Requirements (2/2)

What you will need, cont...

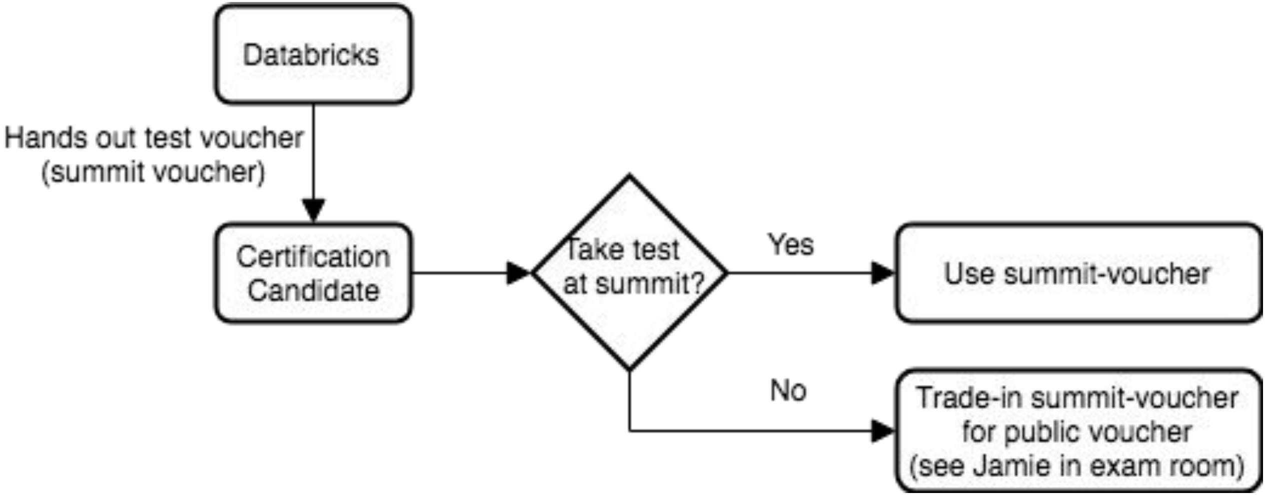
4. We will confirm your identity and proof of purchase
   - Passport
   - Driver’s License
   - Or other State Issued ID

5. A computer will be provided in the test center

6. You have up to 3 hours to complete the exam
   - Take care of your bathroom needs before starting
   - If you leave the testing center during the exam
     we will have to submit your work, completed or not

Test Voucher Code

- Vouchers will be handed out during the morning prep-course.
- Taking the test today - use the handed out voucher (summit voucher)
- Taking the test after the summit, you'll have to trade your voucher for a different voucher (public voucher).


Databricks
Hands out test voucher (summit voucher)
Certification Candidate
Take test at summit?
Yes
Use summit-voucher
No
Trade-in summit-voucher for public voucher (see Jamie in exam room)

Post-Summit Requirements

What you will need

1. High-speed, stable, internet access with open firewalls

2. A webcam (internal or external) and microphone

3. A supported web browser
   - Windows: **Internet Explorer 11 or Edge, Firefox (latest), Chrome (latest)**
   - Mac: **Safari, Firefox (latest), Chrome (latest)**

4. Fore more specifics, see Kryterion's Test Taker Guide https://www.kryteriononline.com/sites/default/files/docs/PreparingForYourExam.pdf

5. For Technical Support email or call olpsupport@kryteriononline.com 1-877-313-2008

Key Points About the Exam (1/3)

- 40 Questions
- Multiple Formats
    - One of Many
    - Many of Many
    - Negated - read carefully!
    - Matching
- Time limit is 3 hours - take your time!

Key Points About the Exam (2/3)

- This is not an open book test
  - Proctors will monitor in-person/onsite tests
  - Kryterion will monitor online tests via camera, mikes and other software

- You must press "submit" in order for the exam to be graded

- Your score by percentage will be displayed immediately

- A passing score is 65%

- If you pass you will receive an email...
  - With an authentic Databricks certificate PDF
  - A unique identification number
  - A digital Databricks Certified Developer Logo

Key Points About the Exam (3/3)

- If you fail, you will have one free retake

- You will need an additional voucher to retake the test
  - See Jamie Murray for the retake voucher
  - This new voucher will not work here at the conference
  - The retake voucher can only be used for remote proctoring or an authorized test center
  - The voucher expires after one year

- See the Kryterion website for authorized testing centers near you or an online proctored exam.

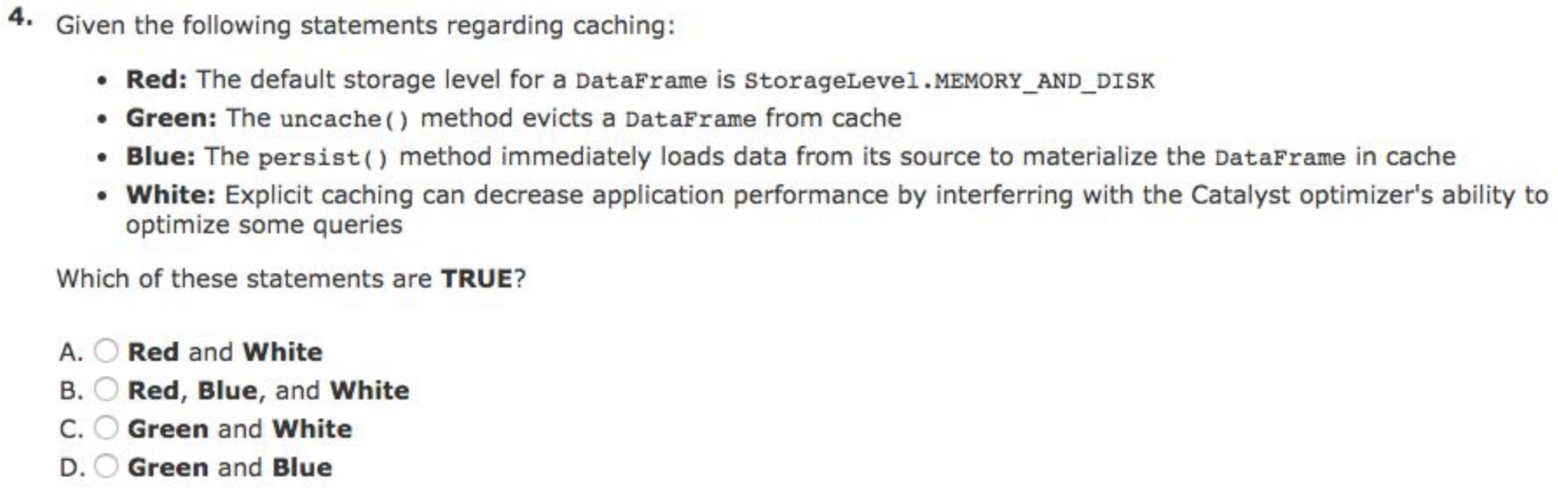
- The Scala and Python exams are functionally identical

Certification Exam Scope

- Understanding breadth of Spark API usage
- Understanding DataFrame API usage
- Applying best practices to avoid runtime issues and performance bottlenecks
- Scaling Spark Applications through API features and architecture
- Integrating Streaming, ML, GraphFrames atop the Spark unified engine
- Solving typical use cases

Format of Exam Questions (1/4)

Select one item that is true or false



4. Given the following statements regarding caching:

- **Red:** The default storage level for a `DataFrame` is `StorageLevel.MEMORY_AND_DISK`
- **Green:** The `uncache()` method evicts a `DataFrame` from cache
- **Blue:** The `persist()` method immediately loads data from its source to materialize the `DataFrame` in cache
- **White:** Explicit caching can decrease application performance by interferring with the Catalyst optimizer's ability to optimize some queries

Which of these statements are **TRUE**?

A. ○ **Red** and **White**

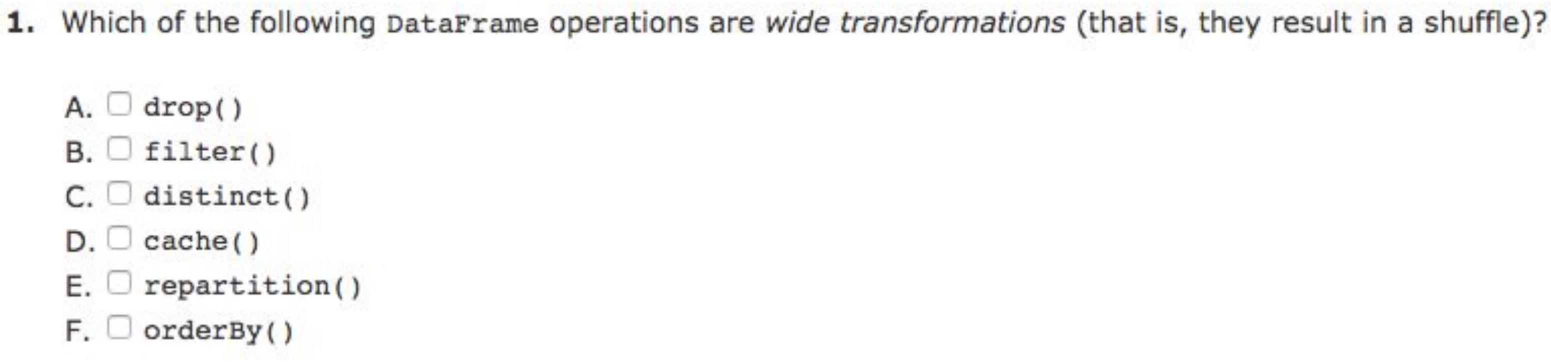
B. ○ **Red**, **Blue**, and **White**

C. ○ **Green** and **White**

D. ○ **Green** and **Blue**

Format of Exam Questions (2/4)

Select multiple items that are true or false



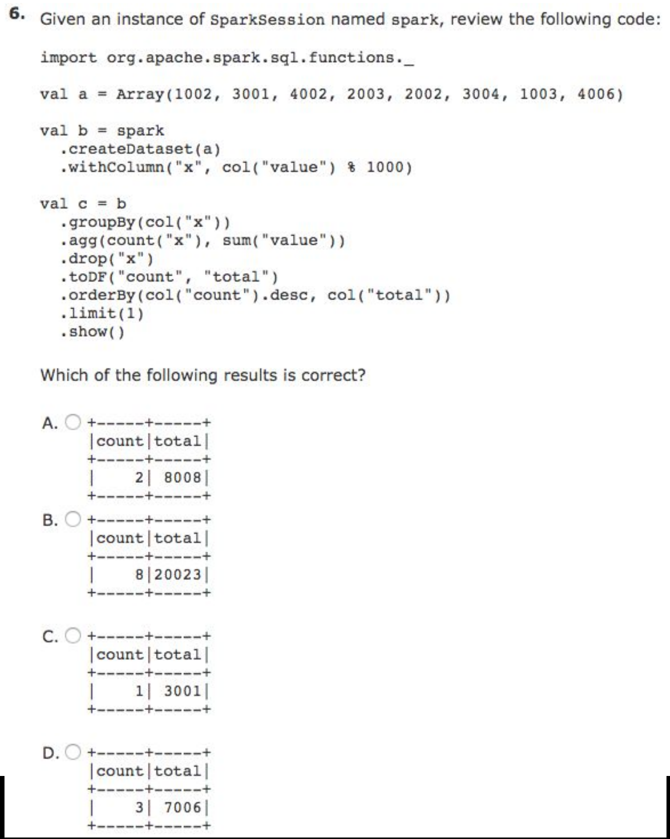
1. Which of the following `DataFrame` operations are *wide transformations* (that is, they result in a shuffle)?

A. □ `drop()`
B. □ `filter()`
C. □ `distinct()`
D. □ `cache()`
E. □ `repartition()`
F. □ `orderBy()`

Format of Exam Questions (3/4)

- Given a code fragment, identify the result(s) it produces

- Given a code fragment, identify errors it contains



6. Given an instance of `SparkSession` named `spark`, review the following code:

```
import org.apache.spark.sql.functions._

val a = Array(1002, 3001, 4002, 2003, 2002, 3004, 1003, 4006)

val b = spark
  .createDataset(a)
  .withColumn("x", col("value") % 1000)

val c = b
  .groupBy(col("x"))
  .agg(count("x"), sum("value"))
  .drop("x")
  .toDF("count", "total")
  .orderBy(col("count").desc, col("total"))
  .limit(1)
  .show()
```

Which of the following results is correct?

A. ○
```
+-----+-----+
|count|total|
+-----+-----+
|    2| 8008|
+-----+-----+
```

B. ○
```
+-----+-----+
|count|total|
+-----+-----+
|    8|20023|
+-----+-----+
```

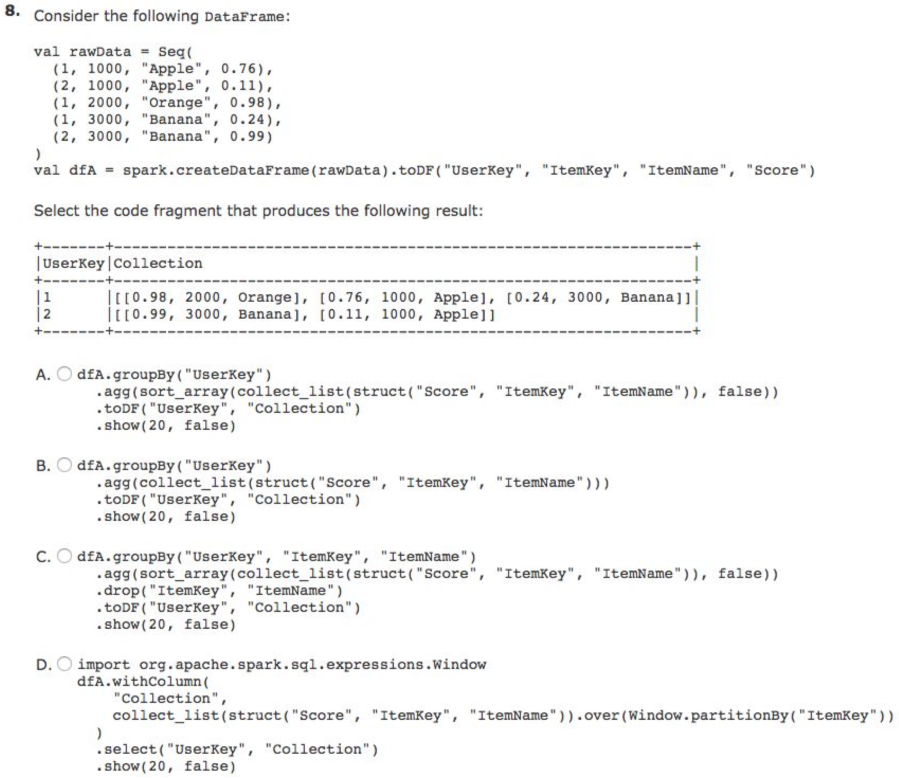
C. ○
```
+-----+-----+
|count|total|
+-----+-----+
|    1| 3001|
+-----+-----+
```

D. ○
```
+-----+-----+
|count|total|
+-----+-----+
|    3| 7006|
+-----+-----+
```

Format of Exam Questions (4/4)

- Given a desired goal, select the code fragment that produces those results

- Given a desired goal, select the design or implementation that minimizes runtime issues or performance bottlenecks



8. Consider the following DataFrame:

```
val rawData = Seq(
  (1, 1000, "Apple", 0.76),
  (2, 1000, "Apple", 0.11),
  (1, 2000, "Orange", 0.98),
  (1, 3000, "Banana", 0.24),
  (2, 3000, "Banana", 0.99)
)
val dfA = spark.createDataFrame(rawData).toDF("UserKey", "ItemKey", "ItemName", "Score")
```

Select the code fragment that produces the following result:

```
+-------+-----------------------------------------------------------------+
|UserKey|Collection                                                       |
+-------+-----------------------------------------------------------------+
|1      |[[0.98, 2000, Orange], [0.76, 1000, Apple], [0.24, 3000, Banana]]|
|2      |[[0.99, 3000, Banana], [0.11, 1000, Apple]]                      |
+-------+-----------------------------------------------------------------+
```

A. ○
```
dfA.groupBy("UserKey")
  .agg(sort_array(collect_list(struct("Score", "ItemKey", "ItemName")), false))
  .toDF("UserKey", "Collection")
  .show(20, false)
```

B. ○
```
dfA.groupBy("UserKey")
  .agg(collect_list(struct("Score", "ItemKey", "ItemName")))
  .toDF("UserKey", "Collection")
  .show(20, false)
```

C. ○
```
dfA.groupBy("UserKey", "ItemKey", "ItemName")
  .agg(sort_array(collect_list(struct("Score", "ItemKey", "ItemName")), false))
  .drop("ItemKey", "ItemName")
  .toDF("UserKey", "Collection")
  .show(20, false)
```

D. ○
```
import org.apache.spark.sql.expressions.Window
dfA.withColumn(
    "Collection",
    collect_list(struct("Score", "ItemKey", "ItemName")).over(Window.partitionBy("ItemKey"))
  )
  .select("UserKey", "Collection")
  .show(20, false)
```

Exam Topic Areas

- 30% - Spark Architecture and Run-time Behavior
- 40% - Spark SQL and DataFrame/DataSet Manipulation
- 10% - RDDs and Low-Level APIs
- 10% - Structured Streaming
- < 5% - Machine Learning
- < 5% - GraphFrames

Spark Architecture and Run-time Behavior

- Spark cluster components and deployment modes

- Caching - cache(), persist(), unpersist(), and storage levels

- Partitioning
  - Initial DataFrame partitioning when reading from data source
  - Repartitioning via coalesce() vs repartition()
  - Controlling number of shuffle partitions

- Performance
  - Catalyst optimizer
  - Identifying performance bottlenecks in Spark applications

Spark SQL and DataFrame/ DataSet Manipulation

- Reading and writing DataFrames

- Transformations, actions, and other operations
  - Wide vs narrow transformations

- Joins
  - Supported Types
  - Broadcast Joins
  - Cross Joins

- Defining and using User Defined Functions (UDFs)

- Window functions

RDDs and Low-Level APIs

- Basic RDD and PairRDD operations
  - Transformations, such as map(), flatMap(), mapValues()
  - Aggregations
  - Actions
  - Joins

- RDD <-> DataFrame conversions

- Accumulator & AccumulatorV2

- Wide Transformations, such as reduceByKey(), groupByKey() etc.

Structured Streaming

- **No** legacy DStream API coverage
- Standard sources and sinks
- Fault tolerance guarantees
- Streaming DataFrame manipulation
  - Aggregation, including using time windows
- Watermarking
- Checkpointing

Machine Learning

- **No** legacy RDD-based APIs coverage

- ML Pipeline basics
  - Initial DataFrame
  - Transformers
  - Estimators

- Model selection
  - Evaluators
  - Parameter grids

- No knowledge about specific algorithms is required

GraphFrames

- **No** GraphX coverage

- Creating a GraphFrame instance

- Basic GraphFrame operations
  - inDegrees(), outDegrees()
  - bfs(), shortestPaths()
  - triangleCount()

- No knowledge about specific algorithms is required

Key API Classes

You should have a strong command of the following classes/functions

- SparkSession
- DataFrame/DataSet
- DataFrameReader and DataFrameWriter
- Column & Row
- org.apache.spark.sql.functions

Spark Study Resources (1/3)

- Instructor-Led Training
- Approx. 75% of exam content
- Databricks' Spark-105
  Apache Spark Programming
- **E-learning Course**
  Getting started with Spark SQL

Instructor-Led Training

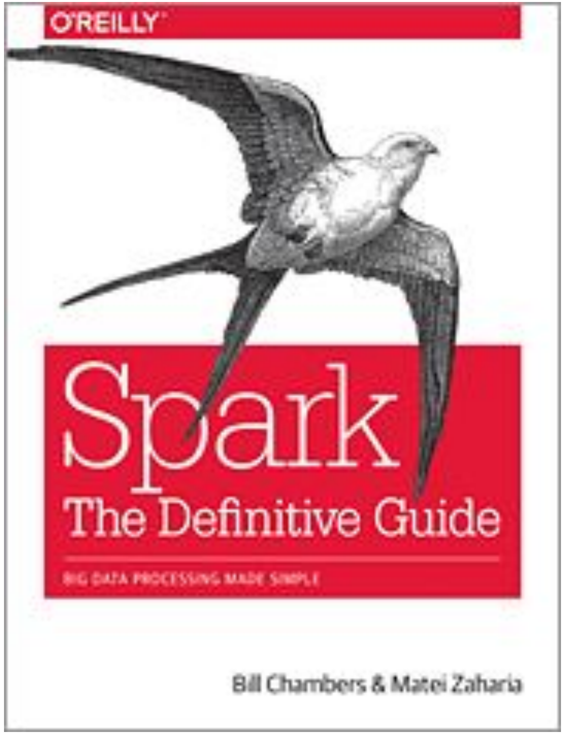
Training from the team that started the Spark research project at UC Berkeley

Our curriculum keeps pace with the platform.

Spark Study Resources (2/3)

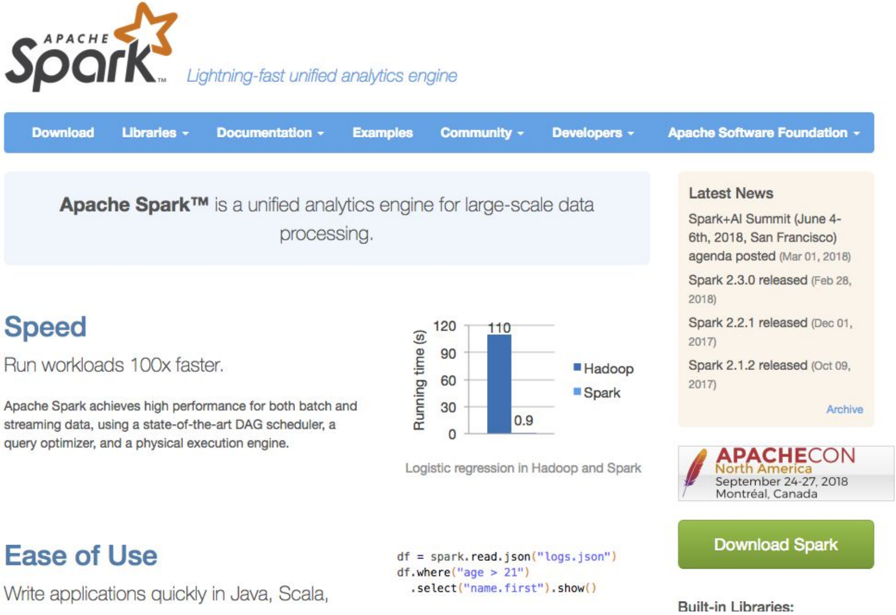
Spark: The Definitive Guide

by Matei Zaharia
and Bill Chambers


O'REILLY
Spark
The Definitive Guide
BIG DATA PROCESSING MADE SIMPLE
Bill Chambers & Matei Zaharia

Spark Study Resources (3/3)

http://spark.apache.org



Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.

Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala,

```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Latest News

Spark+AI Summit (June 4-6th, 2018, San Francisco) agenda posted (Mar 01, 2018)

Spark 2.3.0 released (Feb 28, 2018)

Spark 2.2.1 released (Dec 01, 2017)

Spark 2.1.2 released (Oct 09, 2017)

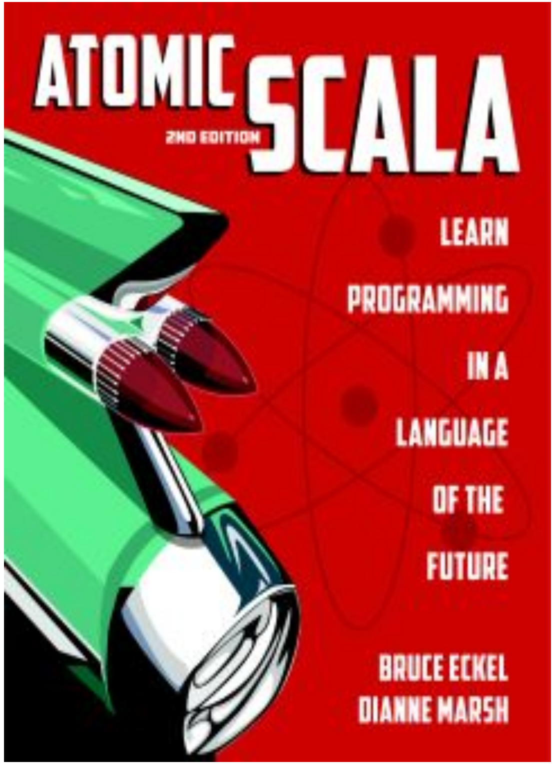
Archive

Built-in Libraries:

Scala Study Resources (1/3)

Atomic Scala,
2nd Edition

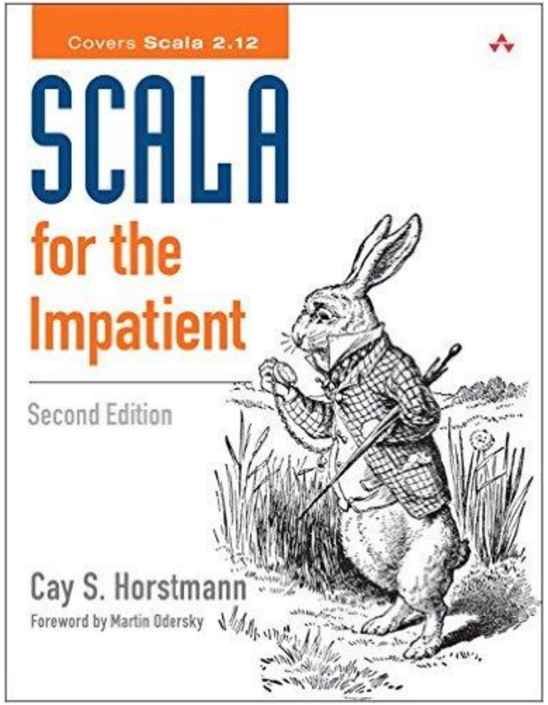
by Bruce Eckel
and Dianne Marsh
http://www.atomicscala.com



Scala Study Resources (2/3)

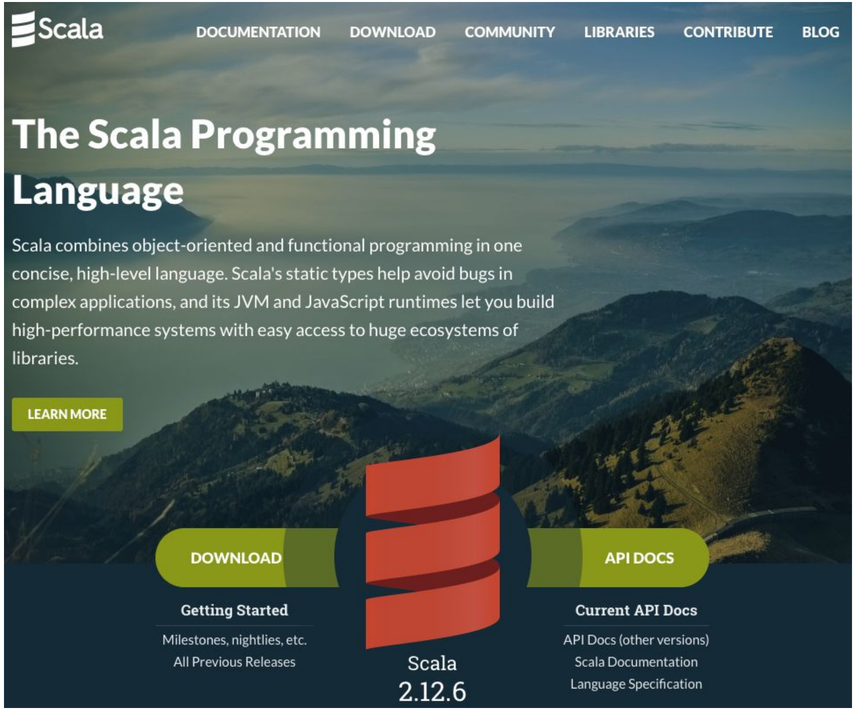
Scala for the Impatient

by Cay Horstmann
https://horstmann.com/scala



Scala Study Resources (3/3)

http://scala-lang.org



The Scala Programming Language

Scala combines object-oriented and functional programming in one concise, high-level language. Scala's static types help avoid bugs in complex applications, and its JVM and JavaScript runtimes let you build high-performance systems with easy access to huge ecosystems of libraries.

Getting Started

Milestones, nightlies, etc.

All Previous Releases

Scala 2.12.6

Current API Docs

API Docs (other versions)

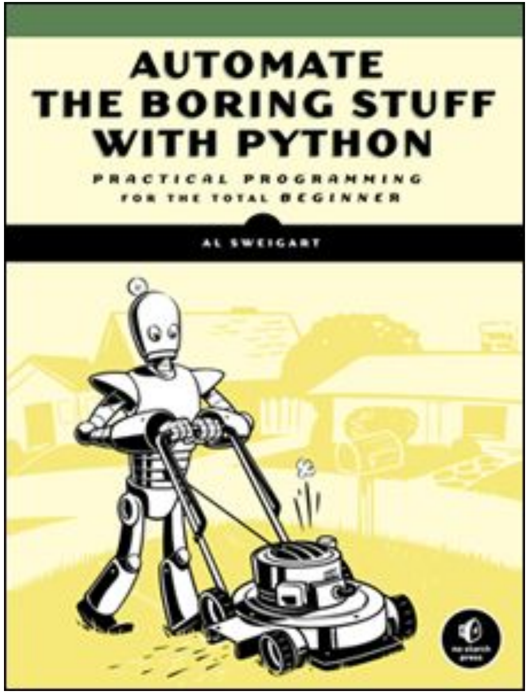
Scala Documentation

Language Specification

Python Study Resources (1/2)

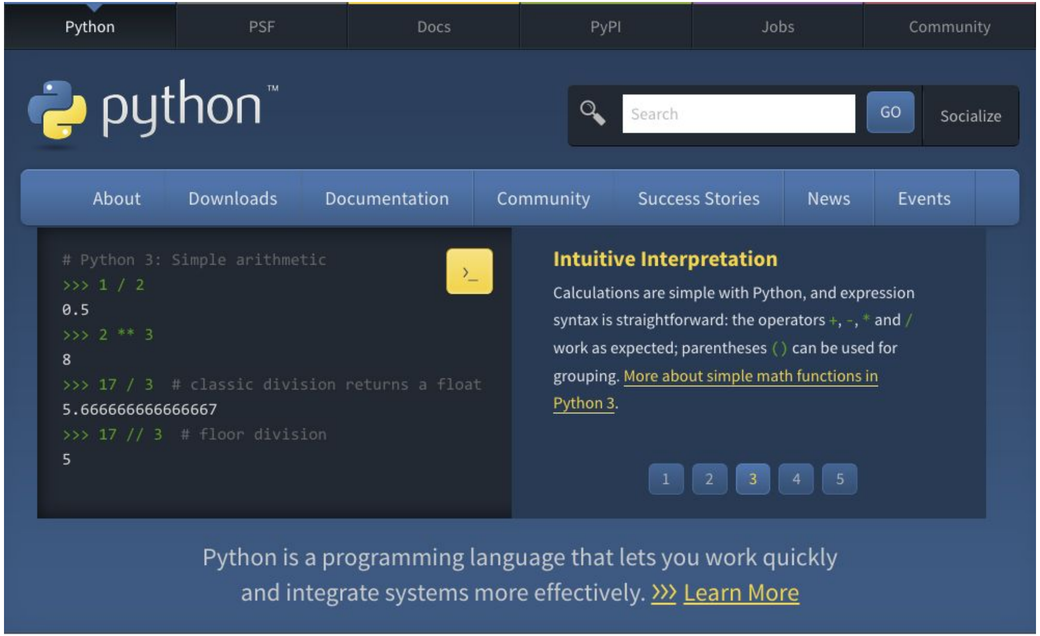
Automate the Boring Stuff with Python

by Al Sweigart
https://automatetheboringstuff.com


AUTOMATE THE BORING STUFF WITH PYTHON
PRACTICAL PROGRAMMING FOR THE TOTAL BEGINNER
AL SWEIGART

Python Study Resources (2/2)

https://www.python.org



```
# Python 3: Simple arithmetic
>>> 1 / 2
0.5
>>> 2 ** 3
8
>>> 17 / 3  # classic division returns a float
5.666666666666667
>>> 17 // 3  # floor division
5
```

Intuitive Interpretation

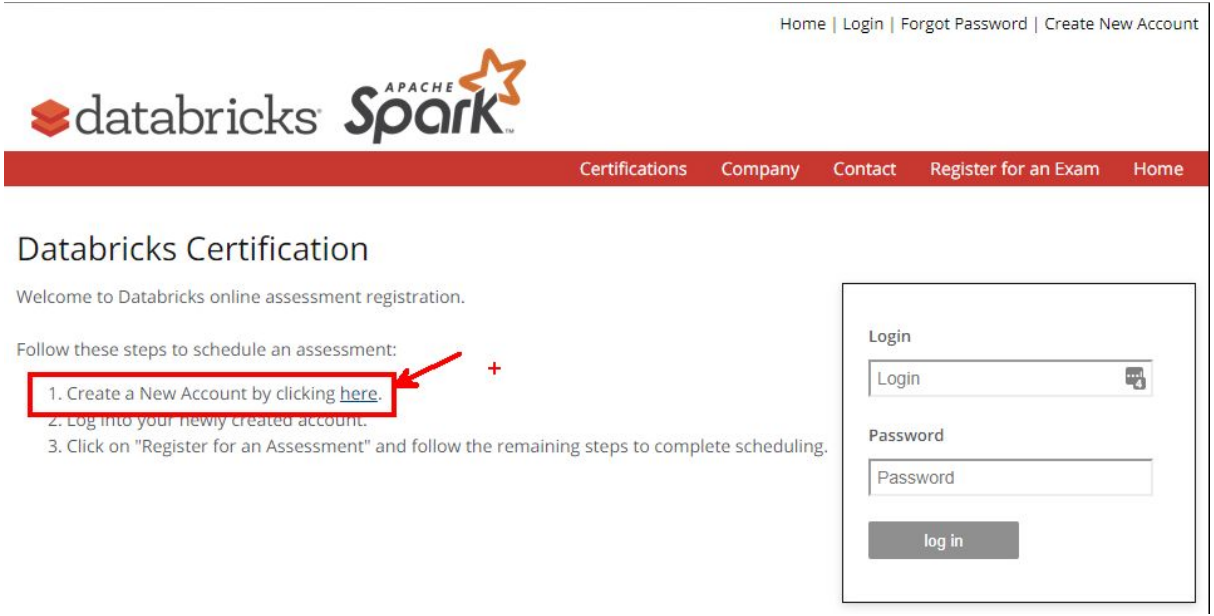
Calculations are simple with Python, and expression syntax is straightforward: the operators +, -, * and / work as expected; parentheses () can be used for grouping. More about simple math functions in Python 3.

Python is a programming language that lets you work quickly and integrate systems more effectively. >>> Learn More

Questions?

Creating An Account (1/3)

1. Go to https://www.webassessor.com/databricks/index.html
2. Select the option to create a new account



Databricks Certification

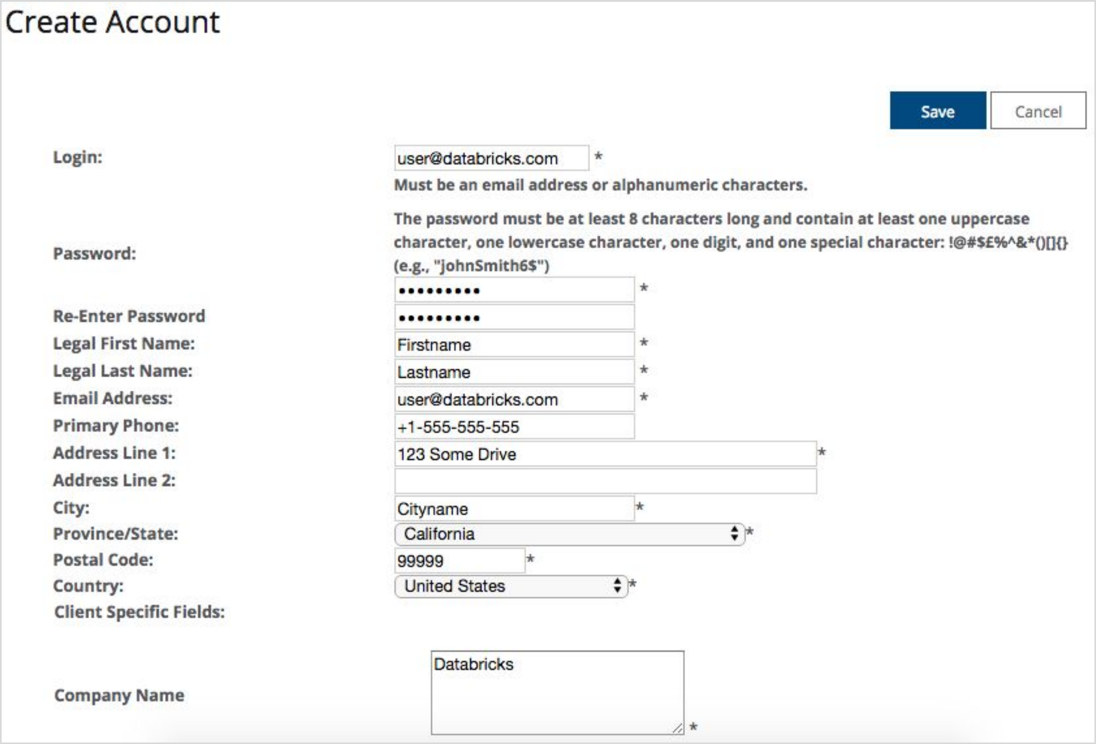
Welcome to Databricks online assessment registration.

Follow these steps to schedule an assessment:

1. Create a New Account by clicking here.
2. Log into your newly created account.
3. Click on "Register for an Assessment" and follow the remaining steps to complete scheduling.

Creating An Account (2/3)

3. Enter your registration information
4. Click **Save**

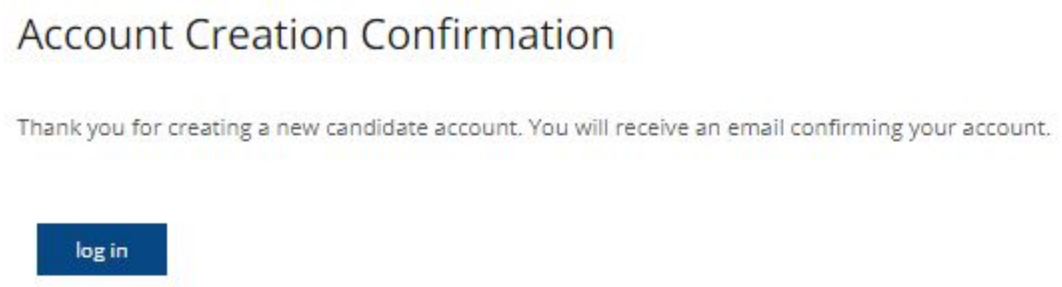


Create Account

Login:	user@databricks.com * Must be an email address or alphanumeric characters.
Password:	The password must be at least 8 characters long and contain at least one uppercase character, one lowercase character, one digit, and one special character: !@#$£%^&*()[]{} (e.g., "johnSmith6$") •••••••••• *
Re-Enter Password	••••••••••
Legal First Name:	Firstname *
Legal Last Name:	Lastname *
Email Address:	user@databricks.com *
Primary Phone:	+1-555-555-555
Address Line 1:	123 Some Drive *
Address Line 2:	
City:	Cityname *
Province/State:	California *
Postal Code:	99999 *
Country:	United States *
Client Specific Fields:	
Company Name	Databricks *

Creating An Account (3/3)

5. Verify Your Email
6. Log in to WebAssessor

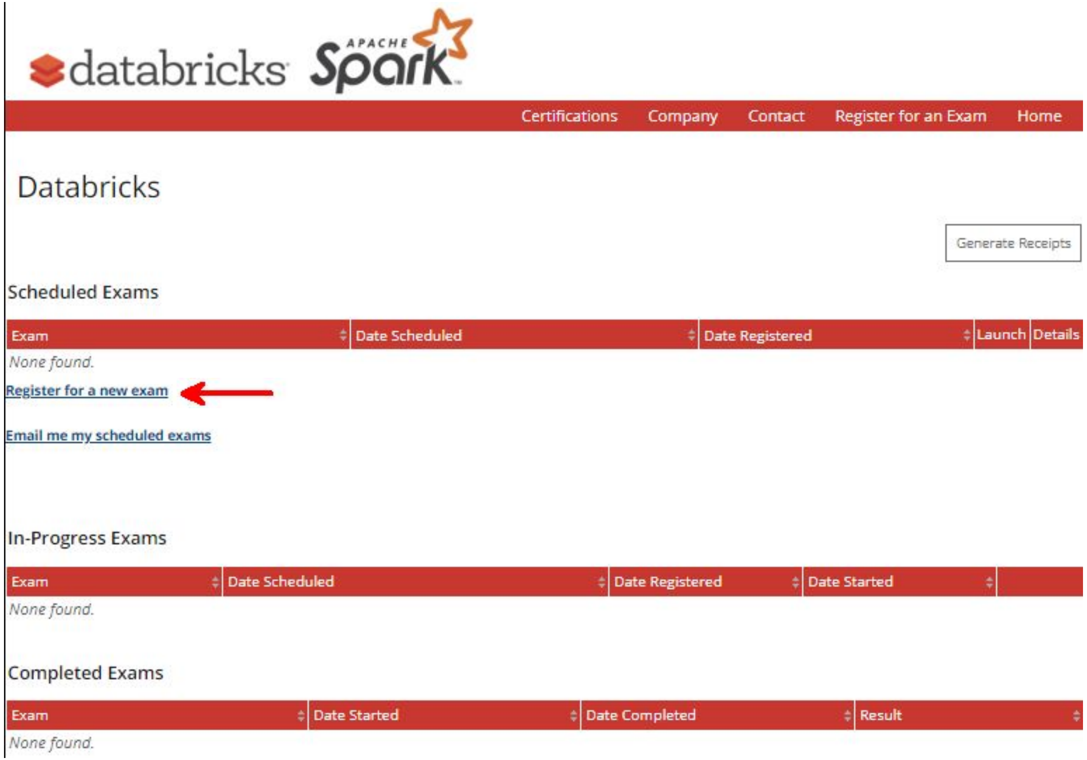


Account Creation Confirmation

Thank you for creating a new candidate account. You will receive an email confirming your account.

Purchase An Exam (1/4)

1. Click the link **Register for a new exam**



Databricks

Scheduled Exams

Exam	Date Scheduled	Date Registered	Launch	Details
None found.				

Register for a new exam

Email me my scheduled exams

In-Progress Exams

Exam	Date Scheduled	Date Registered	Date Started	
None found.				

Completed Exams

Exam	Date Started	Date Completed	Result
None found.			

Purchase An Exam (2/4)

2. Select one of the available exams

The practice/sample exam

Databricks Apache Spark Sample Developer Exam

In The Conference Center

Databricks Certified Developer - Apache Spark 2.x for Scala (Under Private Catalog)

Databricks Certified Developer - Apache Spark 2.x for Python (Under Private Catalog)

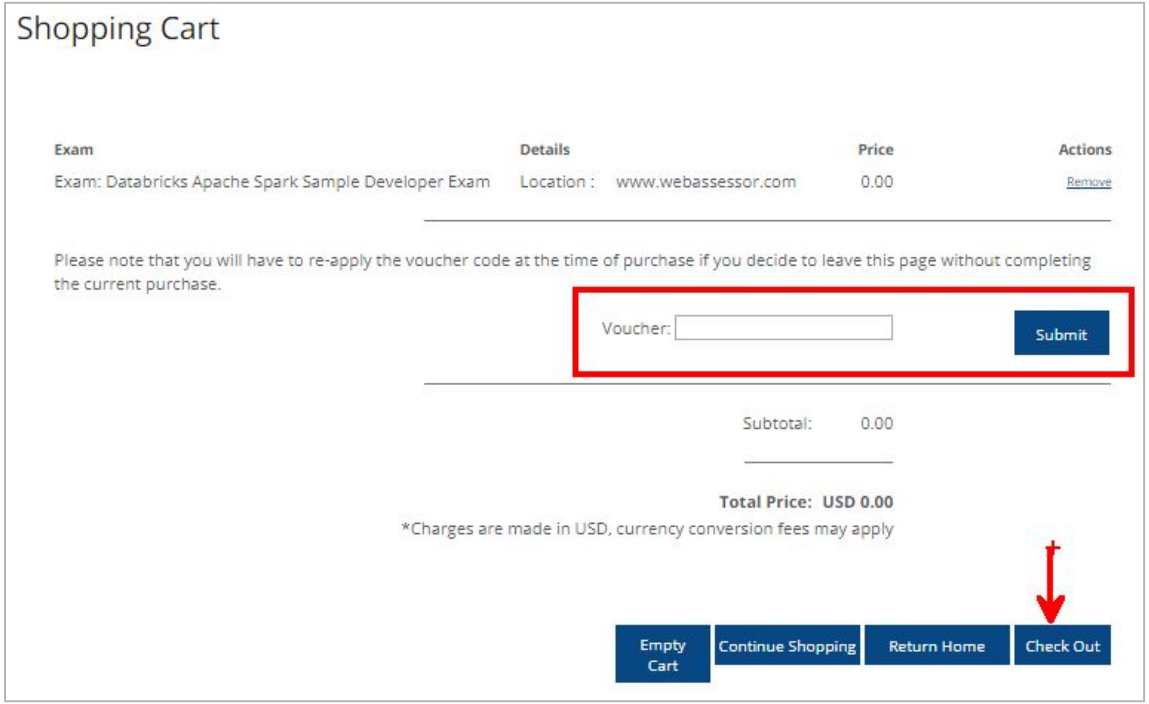
After The Conference

Databricks Certified Developer - Apache Spark 2.x for Scala

Databricks Certified Developer - Apache Spark 2.x for Python

Purchase An Exam (3/4)

3. Enter the voucher and click **Submit** - not needed for our sample exam
4. Click **Check Out** to complete the purchase



Shopping Cart

Exam	Details		Price	Actions
Exam: Databricks Apache Spark Sample Developer Exam	Location :	www.webassessor.com	0.00	Remove

Please note that you will have to re-apply the voucher code at the time of purchase if you decide to leave this page without completing the current purchase.

Voucher:

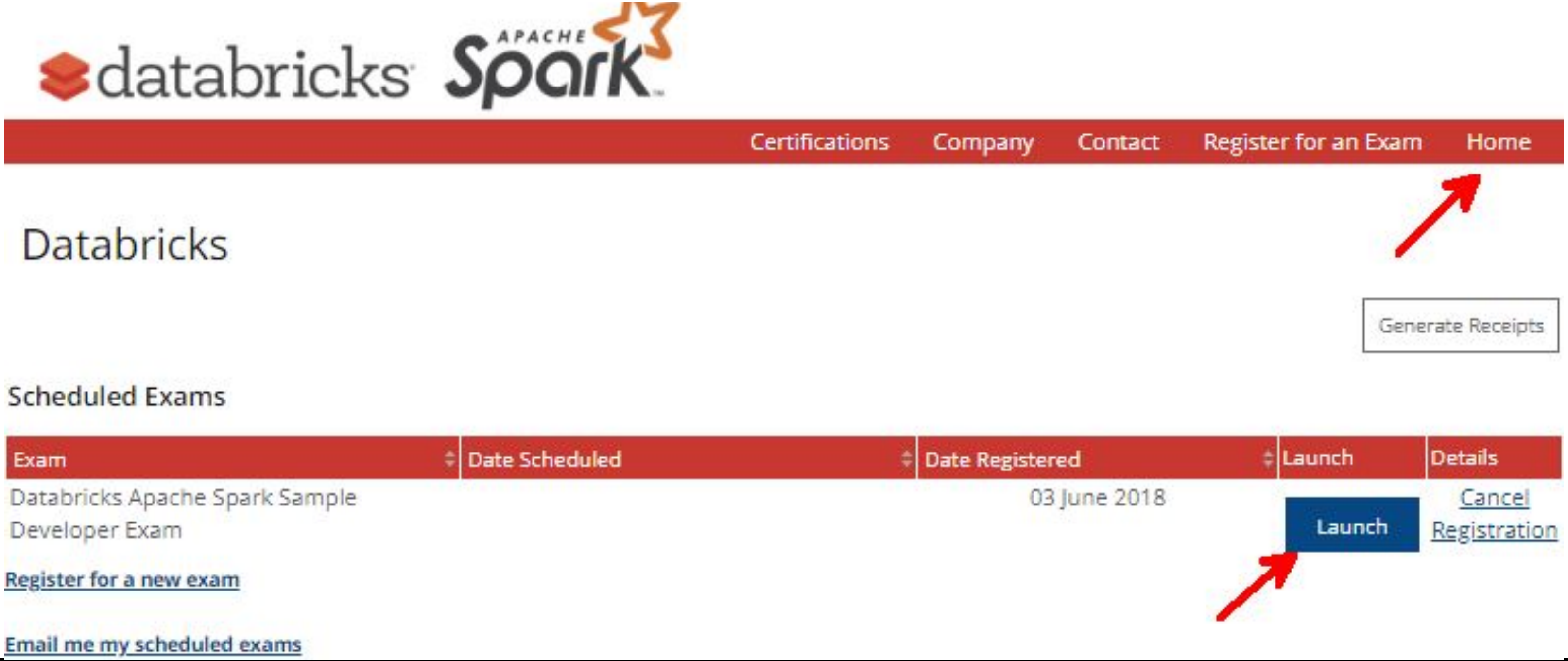
Subtotal: 0.00

Total Price: USD 0.00

*Charges are made in USD, currency conversion fees may apply

Purchase An Exam (4/4)

5. Return to the WebAssessor home page by clicking **Home**
6. Launch the test by clicking **Launch**



Databricks

Scheduled Exams

Exam	Date Scheduled	Date Registered	Launch	Details
Databricks Apache Spark Sample Developer Exam		03 June 2018	Launch	Cancel Registration

Good Luck!

1. The sample exam is not about the questions!

2. Get comfortable with how WebAssessor works

3. Do the sample exam now because the sample exam will not be available tomorrow!

Feedback

https://tinyurl.com/certprep2018