MapR Certified Spark Developer v1

Study Guide

CONTENTS

About MapR Study Guides 3

MapR Certified Spark Developer (MCSD) 3

SECTION 1 – WHAT'S ON THE EXAM? 5
1. Load and Inspect Data 5
2. Build an Apache Spark Application 5
3. Working with Pair RDD 6
4. Working with DataFrames 6
5. Monitoring Spark Applications 7
6. Spark Streaming 7
7. Advanced Spark Programming Machine Learning 8

Sample Questions 9

SECTION 2- PREPARING FOR THE CERTIFICATION 15

Instructor and Virtual Instructor-led Training 15

Online Self-paced Training 16

Videos, Webinars, and Tutorials 18

Blogs & eBooks 19

Datasets 20

SECTION 3 - TAKING THE EXAM 22

Register for the Exam 22

Reserve a Test Session 23

Cancellation & Rescheduling 24

Test System Compatibility 25

Day of the Exam 27

After the Exam - Sharing Your Results 28

Exam Retakes 29

About MapR Study Guides

MapR certification study guides are intended to help you prepare for certification by providing additional study resources, sample questions, and details about how to take the exam. The study guide by itself is not enough to prepare you for the exam. You'll need training, practice, and experience. The study guide will point you in the right direction and help you get ready.

If you use all the resources in this guide, and spend 6-12 months on your own using the software, experimenting with the tools, and practicing the role you are certifying for, you should be well prepared to attempt the exams.

MapR provides practice exams for some of its certifications (coming 2016). Practice exams are a good way to test your knowledge and diagnose the gaps in your knowledge. We recommend that you purchase a practice exam before you commit to a certification exam session.

MapR Certified Spark Developer (MCSD)

The MapR Certified Spark Developer credential is designed for Engineers, Programmers, and Developers who prepare and process large amounts of data using Spark. The certification tests one's ability to use Spark in a production environment; where coding knowledge is tested, we lean toward the use of Scala for our code samples.
Cost: $250
Duration: 2 Hours

The certification has seven sections or domains. These seven sections have specific objectives listed below.

1

What's on the Exam?

Section 1 – What's on the Exam?

The MapR Certified Apache Spark Developer exam is comprised of 7 sections with 33 exam objectives. There are 60-80 questions on the exam. MapR exams are constantly changing and being updated. For that reason, the number of questions on the test vary.

MapR frequently tests new questions on the exam in an unscored manner. This means that you may see test questions on the exam that are not used for scoring your exam, but you will not know which items are live, and which are unscored. These unscored items are being tested for inclusion in future versions of the exam. They do not affect your results.

MapR exams are Pass or Fail. We do not publish the exam cut score because the passing score changes frequently, based on the live test items that are being used.

1. Load and Inspect Data 24%	
1.1	Demonstrate how to create and use Resilient Distributed Datasets (RDDs)
1.2	Apply transformation to RDDs, such as map(), filter(), distinct(), reduceByKey
1.3	Apply actions to RDDs such as count(), collect(), reduce(func), take (n), foreach(func), first()
1.4	Demonstrate how and when to cache intermediate RDDs and persist
1.5	Identify the differences between Actions and Transformations
2. Build an Apache Spark Application 14%	
2.1	Describe how MapReduce jobs are executed and monitored in both MapReduce v.1 and in YARN
2.2	Define the function of SparkContext
2.3	Create a simple Spark application creating a SparkContext with a main method

2.4	Describe the differences between running Spark in the interactive shell vs. building a standalone application
2.5	Run a Spark application in modes including Local, Standalone, YARN, and Mesos
3. Working with Pair RDD 17%	
3.1	Define ways to create Pair RDDs from existing RDDs and from automatically loaded data
3.2	Apply transformations on Pair RDDs including groupByKey and reduceByKey
3.3	Describe the differences between groupByKey and reduceByKey and how each are used
3.4	Use transformations specific to two pPair RDDs, including joins
3.5	Apply actions on pPair RDDs
3.6	Control partitioning across nodes, including specifying the type of partitioner and number of partitions
4. Working with DataFrames 14%	
4.1	Create DataFrames from existing RDDs, data sources, and using reflection
4.2	Demonstrate how to use DataFrame operations including domain methods and SQL
4.3	Demonstrate how to register a user-defined function
4.4	Define how to repartition DataFrames

5. Monitoring Spark Applications 10%	
5.1	Describe components of the Spark execution model such as stages, tasks, and jobs
5.2	Use Spark Web UI to monitor Spark applications
5.3	Debug and tune Spark application logic, as well as Spark configurations, including optimizations, memory cache, and data locality
6. Spark Streaming 10%	
6.1	Describe Apache Spark Streaming architecture components and their purpose
6.2	Demonstrate how to use Spark Streaming to process streaming data
6.3	Demonstrate how to create DStreams using standard RDD operations and stateful operations
6.4	Demonstrate how to trigger computation using output operations such as saveAsTextFiles, foreachRDD, and others
6.5	Demonstrate how to process streaming data using Window operations including countByWindow, reduceByWindow, countByValueAndWindow, and others
6.6	Describe how Spark Streaming achieves fault tolerance

7 .Advanced Spark programming Machine Learning 10%	
7.1	Use accumulators as atomic aggregates and broadcast variables as a way to share data
7.2	Differentiate between supervised and unsupervised machine learning
7.3	Identify use of classification, clustering, and recommendation machine learning libraries
7.4	Describe the basic process for machine learning using Spark

Sample Questions

The following questions represent the kinds of questions you will see on the exam. The answers to these sample questions can be found in the answer key following the sample questions.

1. Which of the following Scala statement would be most appropriate to load the data (sfpd.txt) into an RDD? Assume that SparkContext is available as the variable "sc" and SQLContext as the variable "sqlContext."

   a) `val sfpd=sqlContext.loadText("/path to file/sfpd.txt")`

   b) `val sfpd=sc.loadFile("/path to file/sfpd.txt")`

   c) `val sfpd=sc.textFile("/path to file/sfpd.txt")`

   d) `val sfpd=sc.loadText("/path to file/sfpd.txt")`

2. Given the following lines of code in Scala, identify at which step the input RDD is actually computed.

```
val inputRDD=sc.textFile("/user/user01/data/salesdat.csv")
val elecRD=inputRDD.map(line=>line.split(","))
val elecRDD=inputRDD.filter(line=>line.contains("ELECTRONICS"))
val electCount=elecRDD.count()
```

   a) The inputRDD is computed and data loaded as soon as it is defined

   b) The inputRDD is computed and data loaded when count() is applied

   c) The inputRDD is computed and data loaded when filter() is applied

   d) The inputRDD is computed and data loaded when map() is applied

3. When building a standalone application, you need to create the SparkContext. To do this in Scala, you would include which of the following within the main method?

a)
```
val conf= new SparkConf().setAppName("AuctionsApp")
   val sc= new SparkContext(conf)
```

b)
```
val sc= SparkContext().setAppName("AuctionsApp")
   val conf= SparkConf(sc)
```

c)
```
val sc= new SparkContext()
   val conf= new SparkConf().setAppName("AuctionsApp")
```

4. Which of the following is true of running a Spark application on Hadoop YARN?

a) In Hadoop YARN mode, the RDDs and variables are always in the same memory space

b) Running in Hadoop YARN has the advantage of having multiple users running the Spark interactive shell

c) There are two deploy modes that can be used to launch Spark applications on YARN – client mode and cluster mode

d) Irrespective of the mode, the driver is launched in the client process that submitted the job

5. An existing RDD, **unhcrRDD** contains refugee data from the UNHCR. It contains the following fields: (Country of residence, Country of origin, Year, Number of refugees). Sample data is shown below. Assume that number of refugees is of type Int and all other values are of type String.

   ```
   Array(Array(Afghanistan, Pakistan, 2013, 34), Array(Albania,
   Algeria, 2013, 0), Array(Albania, China, 2013, 12)…).
   ```

   To get the count of all refugees by country of residence, use which of the following in Scala?

   a) val country = unhcrRDD.map(x=>(x(0), x(3))).reduceByKey((a,b)=>a+b)

   b) val country = unhcrRDD.map(x=>(x(0),1)).reduceByKey((a,b)=>a+b)

   c) val country = unhcrRDD.map(x=>x.parallelize())

6. Given the pair RDD **country** that contains tuples of the form **((Country, count))**, which of the following is used to get the country with the lowest number of refugees in Scala?

   a) val low = country.sortByKey().first

   b) val low = country.sortByKey(false).first

   c) val low = country.map(x=>(x._2,x._1)).sortByKey().first

   d) val low = country.map(x=>(x._2,x._1)).sortByKey(false).first

7. There are two datasets on online auctions. D1 has the number of the bids for an auction item. D2 contains the seller rating for an auction item. Not every seller has a seller rating. What would you use to get all auction items with the number of bids count and the seller rating (if the data exists) in Scala?

   a) `D2.join(D1)`

   b) `D1.join(D2)`

   c) `D1.leftOuterJoin(D2)`

   d) `D2.leftOuterJoin(D1)`

8. A DataFrame can be created from an existing RDD. You would create the DataFrame from the existing RDD by inferring the schema using case classes in which case?

   a) If your dataset has more than 22 fields

   b) If all your users are going to need the dataset parsed in the same way

   c) If you have two sets of users who will need the text dataset parsed differently

9. In Scala which of the following would be used to specify a User Defined Function (UDF) that can be used in a SQL statement on Apache Spark DataFrames?

   a) `registerUDF("func name", func def)`

   b) `sqlContext.udf(function definition)`

   c) `udf((arguments)=>{function definition})`

   d) `sqlContext.udf.register("func name", func def")`

Sample Questions Answer Key

1. Option C
2. Option B
3. Option A
4. Option C
5. Option A
6. Option D
7. Option C
8. Option B
9. Option D

2

Preparing for the Certification

Section 2- Preparing for the Certification

MapR provides several ways to prepare for the certification including classroom training, self-paced online training, videos, webinars, blogs, and ebooks.

MapR offers a number of training courses that will help you prepare. We recommend taking the classroom training first, followed by self-paced online training, and then several months of experimentation on your own learning the tools in a real-world environment.

We also provide additional resources in this guide to support your learning. The blogs, whiteboard walkthroughs, and ebooks are excellent supporting material in your efforts to become a Spark Developer.

Instructor and Virtual Instructor-led Training

All courses include:

- Certified MapR Instructor who is an SME in the topic, and is expert in classroom facilitation and course delivery techniques
- Collaboration and assistance for all students on completion of exercises
- Lab exercises, a lab guide, slide guide, job aids as appropriate
- Course Cluster for completing labs provided
- Certification exam fee included – one exam try only, done on the student's own time (not in class)

DEV 3600 – Developing Spark Applications
Duration: 3 days
Cost: $2400
Course Description:

This introductory course enables developers to get started developing big data applications with Apache Spark. In the first part of the course, you will use Spark's interactive shell to load and inspect data. The course then describes the various modes for launching a Spark application. You will then go on to build and launch a standalone Spark application. The concepts are taught using scenarios that also form the basis of hands-on labs.

Dates:
Please see https://www.mapr.com/services/mapr-academy/big-data-hadoop-instructor-led-training

Online Self-paced Training

DEV 360 - Apache Spark Essentials

Duration: 90 minutes
Cost: FREE!
Course Description:

DEV 360 is part 1 in a 2-part course that enables developers to get started developing big data applications with Apache Spark. You will use Spark's interactive shell to load and inspect data. The course then describes the various modes for launching a Spark application. You will then go on to build and launch a standalone Spark application. The concepts are taught using scenarios that also form the basis of hands-on labs.

Lesson 1 – Introduction to Apache Spark

- Describe the features of Apache Spark
- Advantages of Spark
- How Spark fits in with the big data application stack
- How Spark fits in with Hadoop
- Define Apache Spark components

Lesson 2 – Load and Inspect Data in Apache Spark

- Describe different ways of getting data into Spark
- Create and use Resilient Distributed Datasets (RDDs)
- Apply transformation to RDDs
- Use actions on RDDs
- Load and inspect data in RDD
- Cache intermediate RDDs
- Use Spark DataFrames for simple queries
- Load and inspect data in DataFrames

Lesson 3 – Build a Simple Apache Spark Application

- Define the lifecycle of a Spark program
- Define the function of SparkContext
- Create the application
- Define different ways to run a Spark application
- Run your Spark application
- Launch the application

DEV 361 - Build and Monitor Apache Spark Applications

DEV 361 is the second in the Apache Spark series. You will learn to create and modify pair RDDs, perform aggregations, and control the layout of pair RDDs across nodes with data partitioning.

This course also discusses Spark SQL and DataFrames, the programming abstraction of Spark SQL. You will learn the different ways to load data into DataFrames, perform operations on DataFrames using DataFrame functions, actions and language integrated queries, and create and use user-defined functions with DataFrames.

This course also describes the components of the Spark execution model using the Spark Web UI to monitor Spark applications. The concepts are taught using scenarios in Scala that also form the basis of hands-on labs. Lab solutions are provided in Scala and Python. Since this course is a continuation of DEV 360, course lessons begin at lesson 4.

Lesson 4 – Work with PairRDD

- Review loading and exploring data in RDD
- Describe and create Pair RDD
- Control partitioning across nodes

Lesson 5 – Work with DataFrames

- Create DataFrames
- From existing RDD
- From data sources
- Work with data in DataFrames
- Use DataFrame operations
- Use SQL
- Create user-defined functions (UDF)
- UDF used with Scala DSL
- UDF used with SQL
- Repartition DataFrames
- Supplemental Lab: Build a standalone application

Lesson 6 – Monitor Apache Spark Applications

- Describe components of the Spark execution model
- Use Spark Web UI to monitor Spark applications
- Debug and tune Spark applications

Videos, Webinars, and Tutorials

In addition to the classroom and self-paced training courses, we recommend these videos, webinars, and tutorials

1. An Overview of Apache Spark https://www.youtube.com/watch?v=mL5dQ_1gkiA

2. Apache Spark vs. MapReduce – Whiteboard Walkthrough. The following whiteboard walkthrough describes the differences between MapReduce and Apache Spark. https://youtu.be/KzFe4T0PwQ8

3. Parallel and Iterative Processing for Machine Learning Recommendations with Spark https://www.youtube.com/watch?v=Zgbc3Nu2d0o

4. Spark Streaming with HBase https://www.youtube.com/watch?v=O9alig39IKg

5. Adding Complex Data to Spark Stack https://www.youtube.com/watch?v=ls2ytxIOhDU

6. Enterprise-Grade Spark: Leveraging Hadoop for Production Success
   This blog post discusses how you can leverage Hadoop and Apache Spark for production success. https://www.mapr.com/blog/enterprise-grade-apache-spark-leveraging-hadoop-production-success#.VcTiqp1Vikp

7. Getting Started with Spark on MapR Sandbox
   This tutorial shows you how to get started with Spark.
   https://www.mapr.com/products/mapr-sandbox-hadoop/tutorials/spark-tutorial

8. Getting Started with the Spark Web UI
   This post will help you get started using the Apache Spark Web UI to understand how your Spark application is executing on a Hadoop cluster.
   https://www.mapr.com/blog/getting-started-spark-web-ui - .VcP55Z1Vikr

9. Apache Spark: An Engine for Large-Scale Data Processing
   Introduces Spark, explains its place in big data, walks through setup and creation of a Spark application, and explains commonly used actions and operations.
   https://dzone.com/asset/download/35

Blogs & eBooks

We recommend these blog posts and ebooks that can help you prepare for the MapR Certified Spark Administrator exam.

1. The 5-Minute Guide to Understanding the Significance of Apache Spark
   https://www.mapr.com/blog/5-minute-guide-understanding-significance-apache-spark

2. Apache Spark vs. Apache Drill
   https://www.mapr.com/blog/apache-spark-vs-apache-drill

3. Using Python with Apache Spark
   https://www.mapr.com/blog/using-python-apache-spark

4. Getting Started with Apache Spark: From Inception to Production
   https://www.mapr.com/getting-started-apache-spark
   This ebook features guides and tutorials on a wide range of use cases and topics, whiteboard videos, infographics, and more. Start reading now and learn about:
   - What Spark is and isn't
   - How Spark and Hadoop work together
   - How Spark works in production
   - In-depth use cases for Spark (including running code)

Datasets

These are some datasets that we recommend for experimenting with Spark.

1. UCI Machine Learning Repository
   This site has almost 300 datasets of various types and sizes for tasks including classification, regression, clustering, and recommender systems.
   http://archive.ics.uci.edu/ml/

2. Amazon AWS public datasets
   These datasets include the Human Genome Project, the Common Crawl web corpus, Wikipedia data, and Google Books Ngrams. Information on these datasets can be found at http://aws.amazon.com/publicdatasets/

3. Kaggle
   This site includes a collection of datasets used in machine learning competitions run by Kaggle. Areas include classification, regression, ranking, recommender systems, and image analysis. These datasets can be found under the Competitions section at http://www.kaggle.com/competitions

4. KDnuggets
   This site has a detailed list of public datasets, including some of those mentioned earlier. The list is available at http://www.kdnuggets.com/datasets/index.html

5. SF Open Data
   SF OpenData is the central clearinghouse for data published by the City and County of San Francisco and is part of the broader open data program.
   https://data.sfgov.org/data

3

Taking the Exam

Section 3 - Taking the Exam

MapR Certification exams are delivered online using a service from Innovative Exams. A human will proctor your exam. Your proctor will have access to your webcam and desktop during your exam. Once you are logged in for your test session, and your webcam and desktop are shared, your proctor will launch your exam.

This method allows you to take our exams anytime, and anywhere, but you will need a quiet environment where you will remain uninterrupted for up to two hours. You will also need a reliable Internet connection for the entire test session.

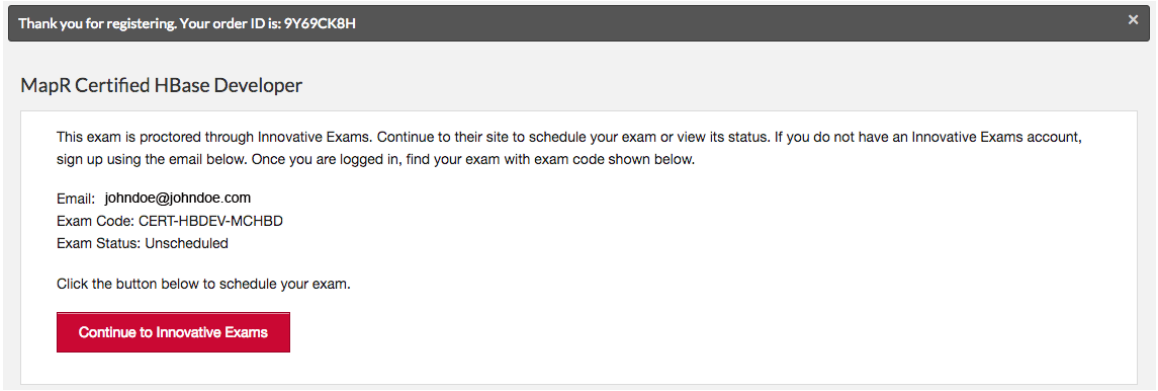
There are five steps in taking your exam:

1) Register for the exam
2) Reserve a test session
3) Test your system compatibility
4) Take the exam
5) Get your results

Register for the Exam

MapR exams are available for purchase exclusively at learn.mapr.com. You have six months to complete your certification after you purchase the exam. After six months have expired, your exam registration will be canceled. There are no refunds for expired certification purchases.

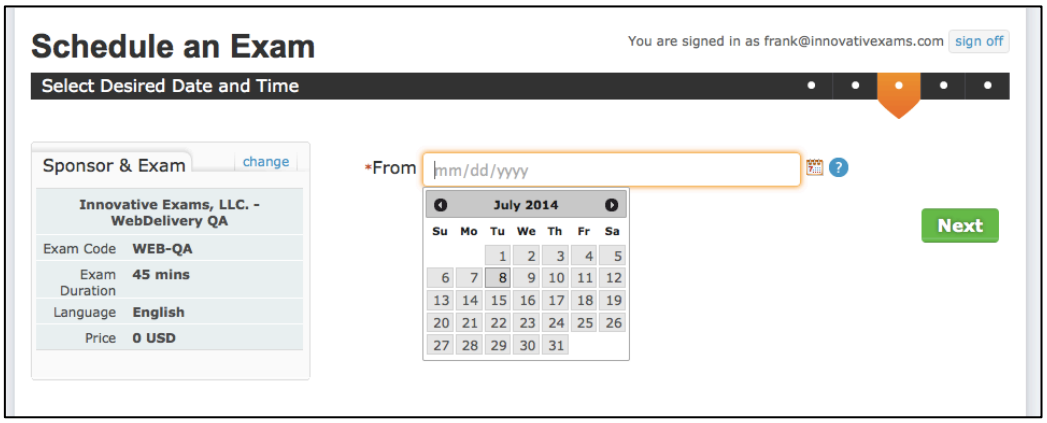
1) Sign in to your profile at learn.mapr.com
2) Find the exam in the learn.mapr.com catalog and click Purchase
3) If you have a voucher code, enter it in the Promotion Code field
4) Use a credit card to pay for the exam
   You may use a Visa, MasterCard, American Express, or Discover credit card. The charge will appear as MAPR TECHNOLOGIES on your credit card statement.
5) Look for a confirmation with your Order ID



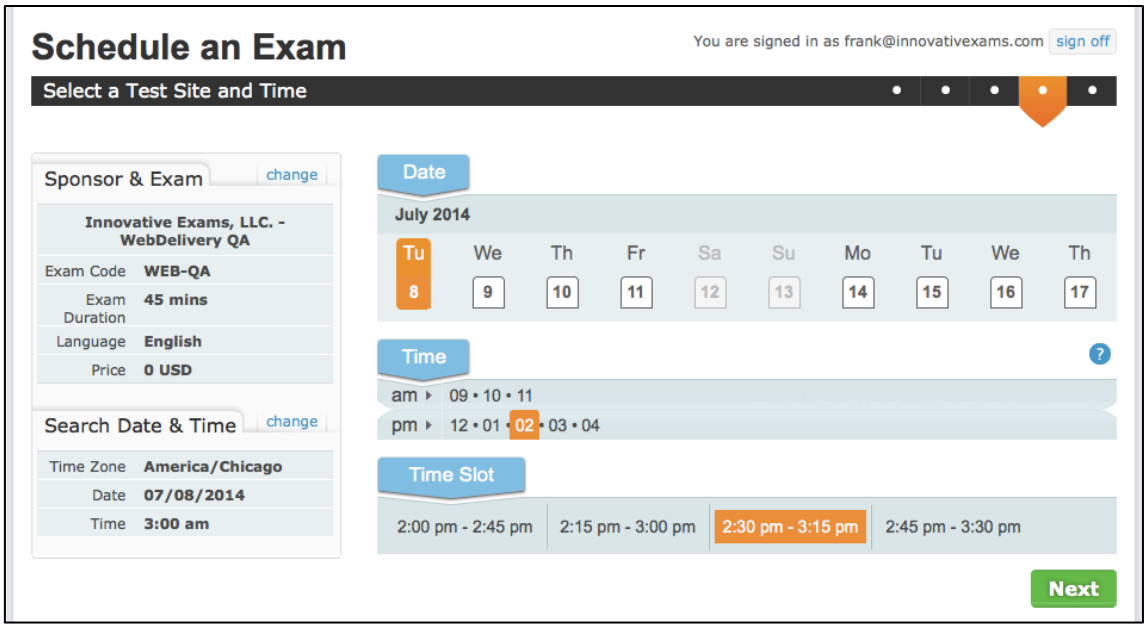
Reserve a Test Session

MapR exams are delivered on a platform called Innovative Exams. When you are ready to schedule your exam, go back to your profile in learn.mapr.com, click on your exam, and click the Continue to Innovative Exams link to proceed to scheduling. This will take you to Examslocal.com.

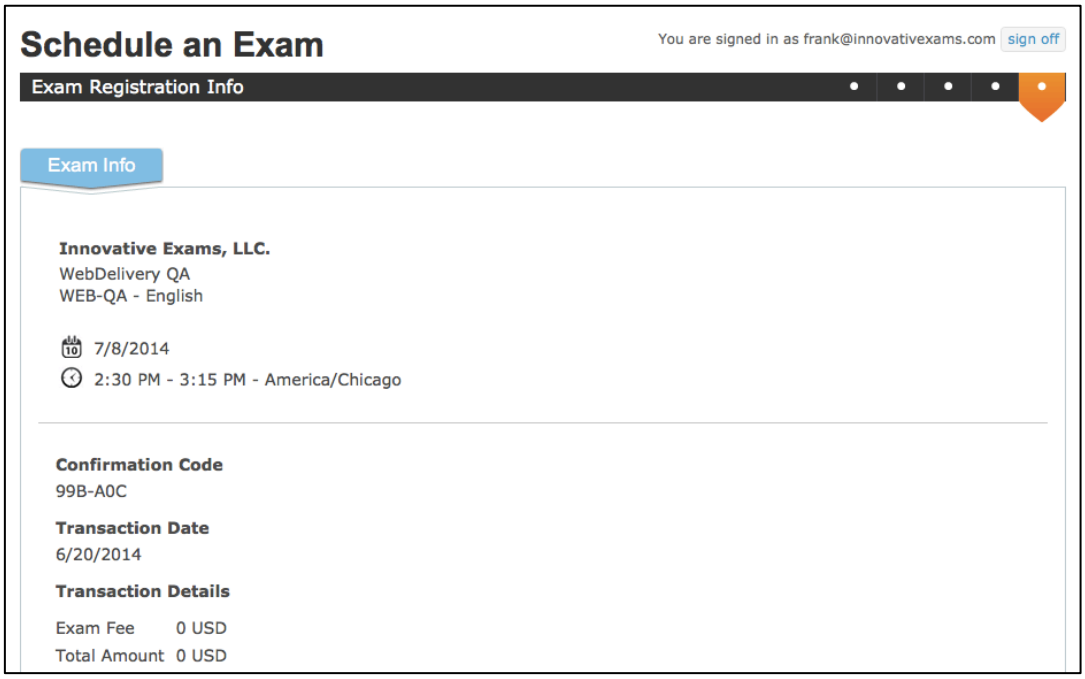
1) Create an account in Examslocal.com
   Make sure to use the same email address that you use in learn.mapr.com
2) Sign in
3) Enter your exam title in the Search field
4) Choose an exam date



5) Choose a time slot at least 24 hours in advance



6) Once confirmed, your reservation will be in your My Exams tab of Innovative Exams

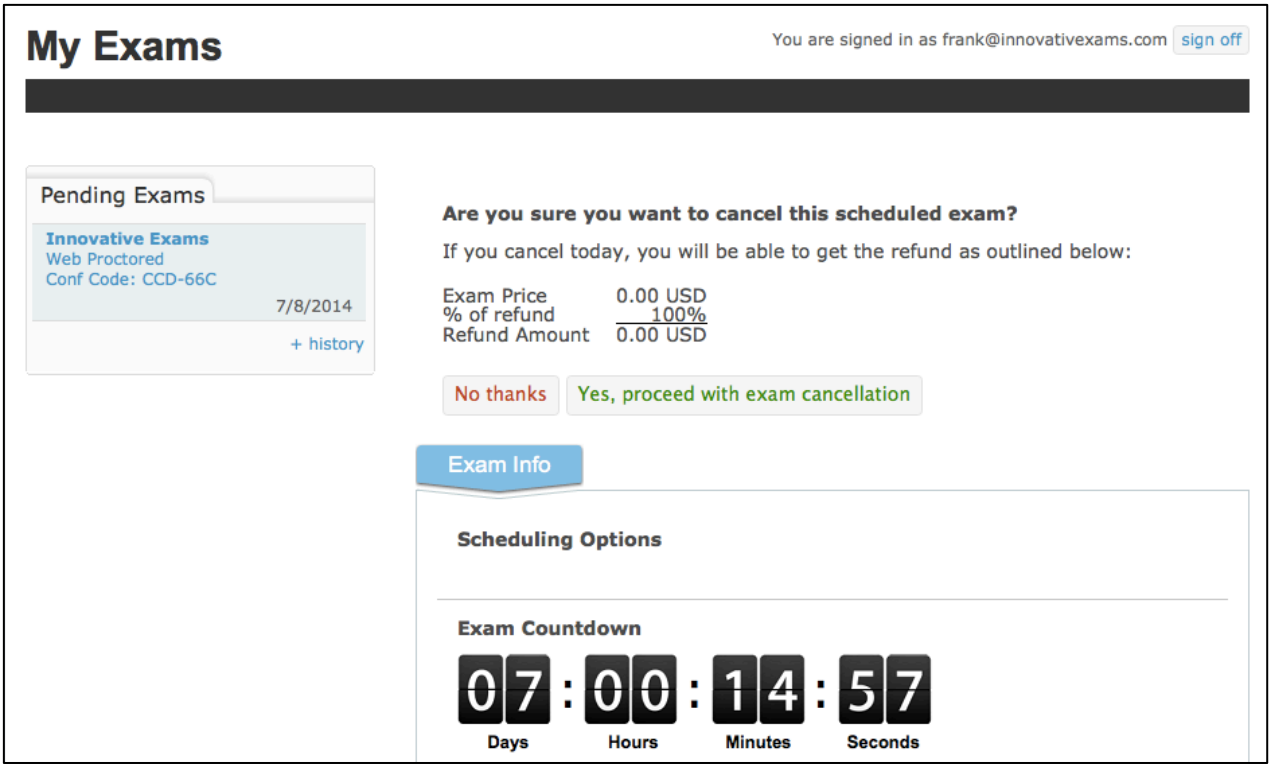


7) Check your email for a reservation confirmation

Cancellation & Rescheduling

Examinees are allowed to cancel or reschedule their exam with 24-hour notice without a cancellation penalty. If they cancel or reschedule within 24 hours of the scheduled appointment, the examinee will forfeit the entire cost of the exam and they will need to pay for it again to reschedule. Examinees must cancel or reschedule their exams more than 24 hours in advance to receive a full refund and remain eligible to take the exam.

To cancel an exam, the examinee logs into www.examslocal.com and clicks My Exams, selects the exam to cancel, and then selects the Cancel button to confirm their cancellation. A cancellation confirmation email will be sent to the examinee following the cancellation.



Test System Compatibility

We recommend that you check your system compatibility several days before your exam to make sure you are ready to go. Go to https://www.examslocal.com/ScheduleExam/Home/CompatibilityCheck

These are the system requirements:

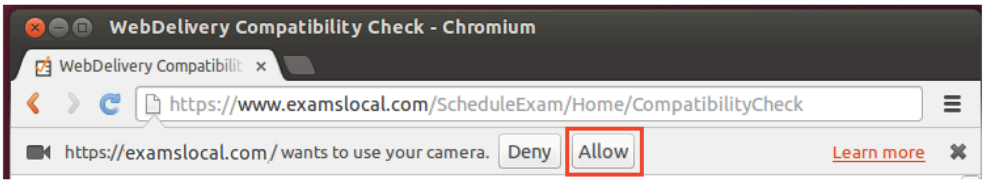
1) Mac, Windows, Linux, or Chrome OS
2) Google Chrome or Chromium version 32 and above
3) Your browser must accept third party cookies for the duration of the exam ONLY
4) Install Innovative Exams Google Chrome Extension
5) TCP: port 80 and 443
6) 1GB RAM & 2GHz dual core processor
7) Minimum 1280 x 800 resolution
8) Sufficient bandwidth to share your screen via the Internet

WebDelivery Compatibility Check

sign in | create an account

Online Proctored Exam Compatibility

Step 1: To enable access to your webcam in Google Chrome click allow in the following prompt.



Step 2: To enable screen sharing in Google Chrome please install the Innovative Exams Extension HERE

Once you have installed the extension please refresh this page.

Step 3: Verify the following minimum requirements

Component	Minimum Requirement	Status	Action Required
Operating System	▪ Windows XP, Vista, 7, 8 ▪ Mac OS X and above ▪ Linux	✓	
Web Browser	▪ Google Chrome or Chromium version 32 and above ▪ Must be able to accept all cookies from third parties	✗	Google Chrome or Chromium required.
Web Camera	Minimum VGA 640 x 480 resolution	✗	Please accept the prompt to allow the use of your camera and microphone.
Innovative Exams Google Chrome Extension	Install Chrome Extension	✗	To enable screen sharing in Google Chrome please install the Innovative Exams Extension HERE Once you have installed the extension please refresh this page.
Ports	**TCP:** port 80 and 443	✗	
Bandwidth	Minimum 500kb/s download and 256kb/s upload	✓	
Hardware Requirements	▪ 1GB RAM & 2GHz dual core processor ▪ Minimum 1280 x 800 resolution		Please ensure that you meet the minimum hardware requirements, listed to the left, before launching your exam at it's scheduled time.
Testing Environment	▪ Room must be quiet, private and well lit. Public spaces such as coffee shops, stores, etc. are not allowed. ▪ Desk must clear of all notes and electronics. ▪ Examinee should reside in the center of the camera frame ▪ No bright lights or windows behind the examinee		Please ensure that you meet the testing environment requirements, listed to the left, before launching your exam at it's scheduled time.

Day of the Exam

- Make sure your Internet connection is strong and stable
- Make sure you are in a quiet, well-lit room without distractions
- Clear the room - you must be alone when taking your exam
- No breaks are allowed during the exam; use the bathroom before you log in
- Clear your desk of any materials, notebooks, and mobile devices
- Silence your mobile and remove it from your desk
- Configure your computer for a single display; multiple displays are not allowed
- Close out of all other applications except for Chrome

We recommend that you sign in 30 minutes in advance of your testing time so that you can communicate with your proctor, and get completely set up well in advance of your test time.

You will be required to share your desktop and your webcam prior to the exam start. YOUR EXAM SESSION WILL BE RECORDED. If the Proctor senses any misconduct, your exam will be paused and you will be notified by the proctor of your misconduct.
If your misconduct is not corrected, the Proctor will shut down your exam, resulting in a Fail.

Examples of misconduct and/or misuse of the exam include, but are not limited to, the following:

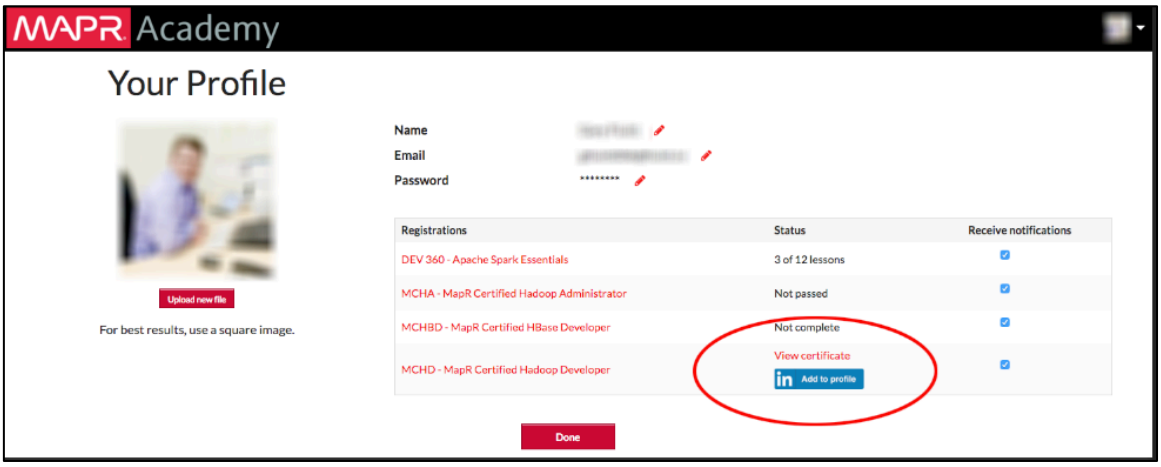
- Impersonating another person
- Accepting assistance or providing assistance to another person
- Disclosure of exam content including, but not limited to, web postings, formal or informal test preparation or discussion groups, or reconstruction through memorization or any other method
- Possession of unauthorized items during the exam. This includes study materials, notes, computers and mobile devices.
- Use of unauthorized materials (including "brain-dump" material and/or unauthorized publication of exam questions with or without answers).
- Making notes of any kind during the exam
- Removing or attempting to remove exam material (in any format)
- Modifying and/or altering the results and/or scoring the report or any other exam record

MapR Certification exam policies can be viewed at: https://www.mapr.com/mapr-certification-policies

After the Exam - Sharing Your Results

When you pass a MapR Certification exam, you will receive a confirmation email from certification@maprtech.com with the details of your success. This will include the title of your certification and details on how you can download your digital certificate, and share your certification on social media.

Your certification will be updated in learn.mapr.com in your profile. From your profile you can view your certificate and share it on LinkedIn.



Your certificate is available as a PDF. You can download and print your certificate from your profile in learn.mapr.com.

Your credential contains a unique Certificate Number and a URL. You can share your credential with anyone who needs to verify your certification.

If you happen to fail the exam, you will automatically qualify for a discounted exam retake voucher. Retakes are $99 USD and can be purchased by contacting certification@maprtech.com. MapR will verify your eligibility and supply you with a special 1-time use discount code which you can apply at the time of purchase.

Exam Retakes

If you fail an exam, you are eligible to purchase and retake the exam in 14 days. Once you have passed the exam, you may not take that version (e.g., v.4.0) of the exam again, but you may take any newer version of the exam (e.g., v.4.1). A test result found to be in violation of the retake policy will result in no credit awarded for the test taken. Violators of these policies may be banned from participation in the MapR Certification Program.