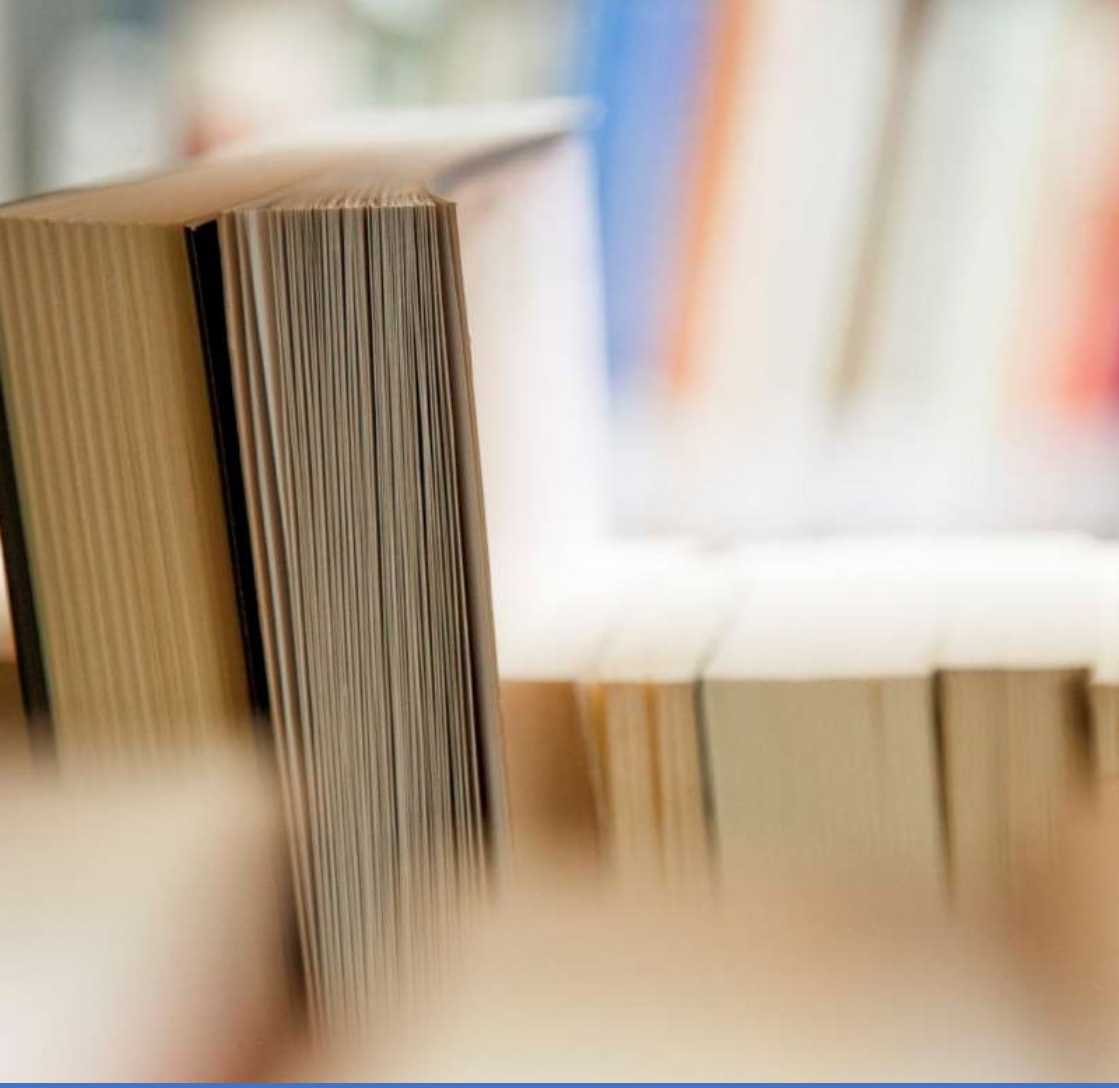


HDPSCD-Hortonworks® Spark Scala Certification Guide

Unofficial, Owned & Prepared by HadoopExam.com

Contents

About book

Apache® Spark is one of the fastest growing technology in BigData computing world. It supports multiple programming languages like Java, Scala, Python and R. Hence, many existing and new framework started to integrate Spark platform as well in their platform e.g. Hadoop, Cassandra, EMR etc. While creating Spark certification material HadoopExam technical team found that there is no proper material and book is available for the Spark (version 2.x) which covers the concepts as well as use of various features and found difficulty in creating the material. Therefore, they decided to create full length book for Spark (HDPSCD Spark Scala Certification) and outcome of that is this book. In this book technical team try to cover both fundamental concepts of Spark 2.x topics which are part of the certification syllabus as well as add as many exercises as possible and in current version we have around 10 hands on exercises added which you can execute on the Hortonworks sandbox, as this book is focused on the Scala version of the certification, hence all the exercises and their solution provided in the Scala. We have divided the entire book in the 7 chapters, as you move ahead chapter by chapter you would be comfortable with the HDPSCD Spark Scala certification. All the exercises given in this book are written using Scala. However, concepts remain same even if you are using different programming language.

Feedback

This is a full-length book from http://hadoopexam.com and we love the feedback so that we can improve the quality of the book. Please send your feedback on hadoopexam@gmail.com or admin@hadoopexam.com

Restrictions

Copyright© Material

online and off-line reviews. Wherever, you use contents make sure full detail of the book is mentioned.

Author had tried as much as his capacity in preparing of this book so that accuracy can be maintained in the presented material. The material sold using this book does not have any warranty or guaranty either express or implied. Neither of the author, publisher, dealer and distributors will be held liable and responsible (explicit/implicit these all parties mentioned are not liable and responsible) for any damages caused or alleged to be caused directly or indirectly by this book. You should note this material as part of your learning process and as time passes material can be outdated and you should wait or look for that latest material.

Author and publisher has endeavoured to provide trademark information about all of the companies and products mentioned in this book. However, we cannot guarantee the accuracy of this information.

Disclaimer:

1. Hortonworks® is a registered trademark of Hortonworks.
2. Cloudera® is a registered trademark of Cloudera Inc
3. Azure® is a registered trademark of Microsoft Inc.
4. Oracle®, Java® are registered trademark of Oracle Inc
5. SAS® is a registered trademark of SAS Inc
6. IBM® is a registered trademark of IBM Inc
7. DataStax ® is a registered trademark of DataStax
8. MapR® is a registered trademark of MapR Inc.
9. Apache® is a registered trademark of Apache Foundation

10. Databricks® is a registered trademark of Databricks Inc

Publication Information

First Version Published: Nov 2019

Edition : 1.0

Piracy

We highly discourage the piracy of copyright material especially it happened online on the internet. Piracy causes the damages to all first of all it damages yourself by not honestly using the correct material, generally pirated material is edited and wrong information is presented which can make big damage as part of your learning process. As well as when you become author and honestly write similar material, piracy will damage your material as well. Hence, don't encourage piracy. If piracy is reduced cost of material will automatically decreases. It also makes damages to author, publisher, dealer and distributors. If you come across any illegal copies of this works in any form on the Internet, then please share the detail with the URL, location or website name immediately on email id hadoopexam@gmail.com we really appreciate your help in protecting author's hard work and also help in reducing the cost of material.

Author/Trainer required

Corporate Trainer: We have many requirements, where our corporate partners need their team to be trained on particular skill sets. If you are already providing corporate trainings for any skills set, then please become our onsite training partner and fill in the form mentioned above and our respective team will contact you soon. You will get very good revenue for sure. However, what we want, you must be able to train our corporate partner resources. What matters to us? Your proficiency in a particular domain/skill and good oral communication skills. You must be able to accessible to learners as well.

Online Trainer: If you are a working professional and master or proficient in any particular skills and feel that, you are capable of giving online virtual trainings e.g. 2 hrs a day until course contents are completed. Please fill in above form and our respective team will contact you or send an email at admin@hadoopexam.com . You will get a very good revenue share for sure. What matters to us? Your proficiency in a particular domain/skill and good oral communication skills. It will certainly not impact your daily work.

Self-Paced Trainings: Ok, you want to work as per your comfortable time and at the same time sharpen your skills. You can consider this option. You can create self-paced trainings on particular domain/skills. Please fill in above form to connect with us as soon as possible. Before somebody else connect with us for the same skill set. Your commitment is very important for us. We respect your work and we will not

sell your work in just $10 to acquire more resources. As we know, it takes a good amount of time and you will provide quality material, so we charge reasonable on that so, you will feel motivated with your work and effort. We respect you and your skill.

Certification Material: You may be already certified professional or preparing for particular certification in a specific domain/skill. So why not use this to make money as well as sharing your effort with other learners globally. Please connect with us by filling form or send email at admin@hadoopexam.com and our respective team will contact you soon.

Author: Yes, we are also looking for authors. Who can write books on a particular technology and what you can get certainly a very good revenue sharing and you can bring the same on your resume or linked in profile to show your excellence? Yes, we are not in need of very good oral communication skills, but good writing skill. However, team will also help you to get work done. Author can be more than one for a particular book. However, we wanted you to be in long relationship. So that you don't just write a single e book, but can create an entire series for a particular domain or skill. Good royalty for sure...

Trending Skills (Not limited these):

Hadoop Spark AWS Cloud Azure Cloud Google Cloud	EMC NetApp VMWare CISCO HP	Adobe Alfresco Apple AppSense AutoDesk	Data Analysis DJango Docker Drupal Graphics	Infrastructre Automation Internet of Things (IOT) ISO Development Java Java Script
JQuery Kali Linux Laravel Linux Machine Learning	Mobile Application Development NodeJS Android Angular JS Arduino	IBM Watson IBM BPM WebMethod Gemfire Liferay	Scala Python Java SQL/PLSQL Ruby	SAP SAS Salesforce Oracle Cloud Redhat

Chapter-1: About HDPSCD Spark Exam

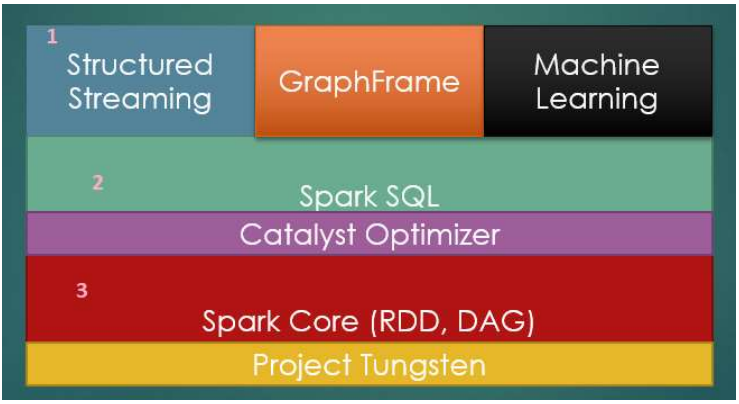
HDPSCD Spark certification exam is conducted by Hortonworks, currently it is evaluated on the HDP-3.x platform, which has Ambari version 2.7. and Spark version 2.3. This certification exam is having 3 sections as below.

1. **Spark Core**: Spark core is the component on which entire Spark framework and other components like SparkSQL, GraphFrame, Machine Learning Library and Structured streaming components are built. Major subject under this section evaluated is Spark RDD API and how to submit the jobs etc.

2. **Spark SQL**: This is one of the biggest changes in Spark 2.x onwards, because Spark 2.x has its own optimization engine known as Catalyst Optimizer. It is highly recommended that if you are using Spark 2.x then use the Spark SQL API and not the RDD API. If you write program using RDD API then you are responsible for optimizing the code that would be executed on the distributed nodes of the Spark cluster. However, in case of Spark SQL, your code goes through various optimization phases before final execution on the distributed nodes. Spark SQL uses the DataFrame and DataSet API which is much simpler and quite intuitive compare to RDD API and if you already have good knowledge of SQL (Structured Query Language) then using Spark SQL is even easier for you.

3. **Spark Structured Streaming**: To process the data in real-time or near real time a completely new framework is developed on the Spark. Previously it was using DStream framework, which is quite complicated for the developer to understand as well as to develop application. Hence, again Spark Development team had created a new Framework for the structured data, this is known as "Structured Streaming". If you can convert your data in well structured (can define a schema) then use the structured streaming which has various in-built feature. One of the well-known features is process exactly once. If you are using any other streaming solution then you would have to write a solution your own to avoid duplicate data processing, in structured streaming this is in-built.

If we take a look on the Spark 2.x components architecture then we can find that each component is built on one another, same exam section is also decided in order.



In above block diagram three components (marked as 1,2 and 3) would be tested or evaluated in the real exam of HDPSCD Spark. This exam is developer oriented as name suggest (Hortonworks Spark Certified Developer). And it is expected from you that you are having some good hands on, with the programming language from one of the below.

- Spark 2.x using Scala
- Spark 2.x using Python

However, it is not expected from you that you are very proficient in any of the programming language. You can consider the trainings provided by "HadoopExam.com" for Scala and Python to have good hands on with these programming languages and good enough to work with the Spark framework. You should be able to write Spark application using one of the programming knowledge. In this exam, it is not expected from you that you have knowledge for both the programming knowledge. Knowing only one programming language either Scala or Python is good enough. Similarly, we have created separate certification preparation material for the HDPSCD using Scala and HDPSCD using Python.

We would be discussing each topic of the syllabus in detail as we move ahead. Syllabus remain same whether you use Python or Scala. However, Spark application can also be written using Java and R language also. But there is no certification available for R and Java language.

HDPSCD Real Exam conditions

1. Total 120 minutes (2 Hrs) for the exam

2. You should be able to score minimum 75% to pass this certification exam
3. Currently exam is conducted in English language (more detail, you can always check on Hortonworks or Cloudera websites)
4. Cost of the exam is $250 (Sometime Cloudera/Hortonworks runs the offer as well for discount).

HDPSCD Exam formatting and structure

This certification exam is completely Hands on exam and there would not be any multiple-choice questions. During the exam usually you would be given 7-10 tasks which you would have to complete. Most of the learners who had solved all 89 problems given by the HadoopExam.com are able to score 100% marks in real exam.

In real exam you would be given multi-node (or single node) HDP cluster which would be running HDP 3.0 platform which would be having Ambari Cluster management software 2.7 version (We have released another book for Ambari Interview Questions) . During the exam **it is not expected** from you to know the Ambari and how to use it to start and stop the cluster components. It is good enough if you know the following components.

1. How to access Spark 2.x platform on the HDP
2. How to access data stored in Hive
3. How to read and write data on the HDFS using HDFS components

4. You should be comfortable to use “vi” editor and some basic commands.

Suppose you have been given 7 tasks then you have to complete at least 5 tasks to successfully pass the certification exam. As per HadoopExam.com experience and previous learners feedback the task given during the exam are not that complicated, but rather much easier compare to your real time projects. The major factor which contribute in your real exam is, how do you manage time and on average within 10 to 15 mins, you should be able to complete each individual exercise. However, we have seen that many of our learners spend more time on some relatively complex exercise and much lesser time on the simple exercises. **Universally known Hint: Always attempt simpler exercise first to pass the exam**. You must preserve the code you have written for solving the exercise and result should be saved on the given location. Because evaluator can check your partially solved problem as well, and it depend on his discretion to give marks on that partially solved problem

How possibly HDPSCD exam can be evaluated:

As this is a hands-on exam and not a multiple choice, evaluators have to put some effort on examining your solution. Sometime you may not be able to achieve desired result, and you might have completed more than 90% of task. Then evaluators must consider your effort as well. So, evaluators don’t want to take any decision in hurry and they want exam to be evaluated properly, without doing any

injustice to you. And Hortonworks university takes around few business days (Almost 2 weeks or less) to report or announce your result. As it should be done by Hortonworks university only. It cannot be done by any training vendor which had partnered with Hortonworks or proctors who monitor your exam.

In your final result Hortonworks would be reporting how much you have scored and what is the desired passing score.

Chapter-2: Individual Task and Assessment description

HDPSCD exam number of tasks and exam pattern

It depends on how complex problem tasks would be given to you and based on that number of tasks would be decided. Usually there are around 7 to 10 tasks would be given. In HadoopExam.com certification preparation material we have covered almost all possible questions and answer pattern. So, you must go through that practice material before appearing in your real exam. In practice material as well, you would be given a particular task, which may or may not depend on the previous task. Also, you can download all the required data to solve that particular problem, and step by step solution would be given to that problem task. Soon videos for the selected problem and answer would be provided in the practice material. Towards the end of this book we would be discussing around 10 problem tasks from HadoopExam.com practice material.

In real exam also may be possible that your next tasks depend on another task. But most of the cases we have seen that, you are given independent tasks. If dependent tasks given then you could face situation something like below.

Example Task-1: You would be given data, which can be mostly in the following format.

- JSON
- CSV (most of the time learners are getting questions with JSON and CSV data)
- Parquet
- Avro
- Text
- ORC (Apache Hive related problem would be related with this format)

You would be given all the data on the HDFS. So, you should be having experience how to access the data using HDFS command line. In your given tasks you should write a program such that you read data from the HDFS file system using Spark API, once data is loaded in RDD or DataFrame you need to apply the filters on that data in some sorted order (ascending or descending) and save the final result in HDFS file system.

You would be given HDP platform, so you must not have to do any specific settings to read and write data on HDFS. All the settings are pre-done and Spark by default read data from the HDFS file system only. So, it is highly recommended that to complete your task, you are able to persist your final result or data on the given location on the HDFS in the desired format like JSON, CSV, AVRO, Parquet, Orc etc.

Other possible variation is, while writing final output, you would be asked to write entire data in a single file or in n-

number of files. So, for different categories or dates write data in respective directory on HDFS platform. Again, you should be aware, how to write data in a single file or multiple n-number of files.

Example based on which category would be created are

1. Save all data related to books category must go in same directory.
2. Save all data related to electronic equipment should be saved in separate file.
3. Save all fitness product data in another file.

Important notes:

- Don't forget to save the code you have written.
- Don't forget to save the result
- Be careful you don't delete the files, data and solution created by yourself.

Example Task-2: It may be possible your next task depends on your previous task. However, this is not always the case. It may be possible that in your first task Hortonworks may give you question in CSV format and next question they would give you in JSON data. Hence, it is important you know how to deal with the data in various format. And very similar you have to apply filter, sorting, aggregations, group by, union etc operations with the Spark SQL and RDD API.