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

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 You should note this material as part of your learning process and as time passes material can be outdated and you should wait or look for that latest material.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	EMC	Adobe	Data Analysis	Infrastructre Automation
Spark	NetApp	Alfresco	DJango	Internet of Things (IOT)
AWS Cloud	VMWare	Apple	Docker	ISO Development
Azure Cloud	CISCO	AppSense	Drupal	Java

QuickTechie

Google Cloud	HP	AutoDesk	Graphics	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

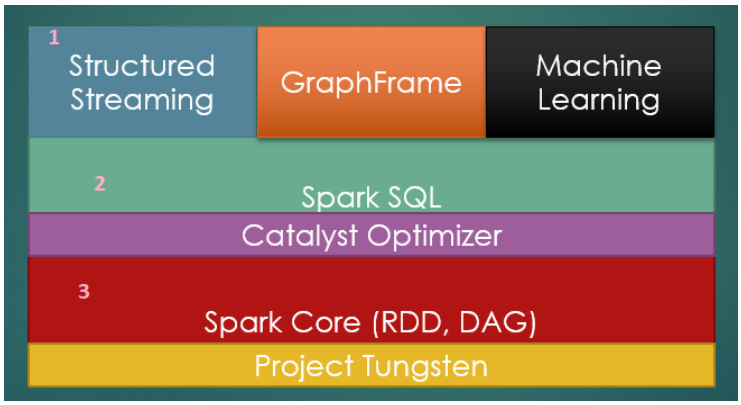
HDPSCD Spark certification exam is conducted by Hortonworks, currently it is evaluated on the HDP-3.x platform, which has Ambari version 2.7. and Spark version 2.3. This certification exam is having 3 sections as below.

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.

Example Task-3: Next possible question they might ask you to use SparkSQL explicitly by writing the SQL query get the desired result. Most likely using Hive Query Language to save or extract data in Hive table from ORC format. Let's say, you have to

- Load data in one of the given Hive table.
- Extract and read data from Hive table.

- Apply filter, aggregation, sorting on the data
- Save the extracted data in some HDFS file.
- Save extracted data in another Hive table.

We have seen, sometime you can get 2-3 tasks based on Apache Hive and ORC data. Number of tasks depend on the complexity of the exam. And you have to write and read data from Sequence, ORC, Parquet, JSON, CSV, text and Avro file.

Example Task-4: In this case you would be given task where you have to generate final result using some analytics SQL query. Hence, it is expected from you to be able to write some SQL analytics query to get the result like count, max, min, standard deviations etc. This type of tasks may include things like below

- Join the data
- Union two or more datasets
- Subtract one dataset from another dataset
- Apply filter conditions on the data
- Rename particular column in the dataset
- Add new column
- Drop an existing column

Example Task-5: It is most likely that you would be getting one of the tasks from below components.

1. Broadcast variable
2. Accumulator variable

Broadcast variable: Broadcast variable task mostly related to join the data locally on each node by distributing smaller dataset on all the nodes in the HDP cluster. So, learn how to use the broadcast variable. Many of time user tries to broadcast RDD directly which is not possible. You have to first convert it into list and then using the SparkContext, broadcast the list.

Accumulator: Accumulators are very similar to counter in the Hadoop distributed framework. This is helpful when you need to get some aggregated data from all the nodes in the cluster to the driver nodes.

To have a good practice with the Accumulator and Broadcast variable go through the all 89 practice questions provided by HadoopExam.com .

Example Task-6: You would be getting tasks related to Data Engineering pipeline which may require following things to be done

- Extract common columns from various datasets
- Process data to join using common column
- Create small ETL pipeline
- Apply analytical function to query the data

It is all up to you how you solve such problem either using Spark SQL query or using Hive QL.

Example Task-7: You would already been given a Spark application stored somewhere locally on the HDP platform. And Hortonworks wants you to submit that application on HDP cluster using YARN (Yet another resource negotiator) scheduler. This is relatively simple task. We don't see that you have been asked to write entire Spark application using Scala (this require lot of time and knowledge which is not related to Hadoop and Spark).

Example Task-8: With the recent update on the syllabus Hortonworks added structured streaming. So certainly, you would get one exercise based on this topic. In this exercise possibly they can ask you to write one streaming application, which can read data from

- Socket
- From HDFS filesystem
- Rate System

And apply some transformation in real time on the streamed data and finally save the result. And during the data stream they can ask you to apply some settings like

- Set processing time
- Ser watermark on the stream
- Set trigger interval
- Set slicing window

However, Hortonworks had not mentioned this all topic explicitly in their syllabus. But these are very relevant with the Streaming application and features of the Spark structured streaming.

Overall syllabus, if you check there are 22 different sub topics are mentioned in the syllabus. And to successfully pass the exam, please prepare at least 95% of all these subtopics (around 20 to 21 topics at least). In sometime we would be discussing all 22 subtopics of the syllabus.

Chapter-3: Dissection of the HDPSCD Spark Scala exam

RDD vs SparkSQL DataFrame API

Many of our learners are getting confused with the RDD API is being part of the syllabus. And as a programmer it is always recommended to you by the Spark community that you should avoid using the RDD in your program if you are already on the Spark 2.x or later version and should use the SparkSQL API or DataFrame/Dataset API. Why?

Yes, that is true as much as possible you should avoid using RDD in your program, if you are already using Spark 2.x version. However, for the exam it is fine if you write code using RDD API to solve your problem task. Hortonworks want you to solve the problem and reach to the desired result using Spark framework. It really does not matter whether you use RDD API or Spark SQL API.

HadoopExam.com also recommends that you know the basics of the RDD and RDD API. However, try to avoid using RDD API if possible, in your program until and unless it is absolutely necessary, for example with the broadcast variable and accumulator you have to use this.

Why you should not use RDD API in your real time project

- RDD API is relatively complex then Spark SQL API
- RDD API does not uses the newly built catalyst optimizer. Which is developed for optimizing the Spark SQL code, without the involvement of the programmer.
- SparkSQL converts Scala datatypes in its own type to use custom serialization.
- Spark leverage highly optimized Project Tungsten to use the modern hardware and cache your code instructions as close to CPU as possible. It uses all three low level caches L1, L2 and L3 efficiently.

Submitting your problem solution

You must preserve your code and save in some text file locally on the same host or the directory in which you have been asked during the exam (instruction should be provided for the same, in the exam), which you are using during your real exam (not on your local desktop). All of your program should be written on the spark-shell or pyspark shell. However, the entire output generated for a particular application should be saved in a directory which is specified in the task or problem statement, mostly that needs to be saved on the HDFS. Otherwise it would be considered that you are not able to complete the given task.

Usually most of our learners got 7 questions or tasks to be completed in 2 hours. And out of 7 tasks you must be able to complete 5 tasks to pass this certification exam. During the exam you would not be provided any IDE (Integrated Development Environment e.g. eclipse, IntelliJ etc).

How to practice for HDPSCD Certification Exam

First of all, you should have some environment to practice all 89 questions or assignment provided by HadoopExam.com and become Spark 2.x expert. You should avoid copying and pasting the code from HadoopExam practice material. Rather write entire solution your own. So that you would

become expert and comfortable with the Spark API and quickly solve the exercise in real exam. There are two possible option for creating practice environment

1. Setup Hortonworks sandbox on your local computer. However, for that your computer must have 16 GB RAM
2. Setup Hortonworks sandbox in the cloud environment like Azure, AWS or on Google cloud. You would be charged for that as per the respective cloud provider fees, that is based on the how long you use their environment.

HadoopExam.com soon release the videos to setup these practice environments and you can follow the instructions to setup the same. All the practice questions provided on the HadoopExam are tested and executed on single node HDP cluster with Spark 2.x platform. In future we would be adding videos as well for selected exercises or for the problem tasks. You must practice as much as possible on the Hortonworks sandbox. By practicing with the Hortonworks sandbox you would be comfortable with the

- Starting and stopping spark-shell or pyspark shell
- You should be able to run Hive commands
- You should be able to run HDFS commands
- You should know, how to submit Spark jobs.
- Setting up various properties in Spark shell

Difficulty level of the real exam

From our learners' feedback we come to know that this exam is not very difficult. Also, the task you would be performing are not administratively complex. If you have been working for 3-4 months on Spark and well-practiced all the 89 tasks or questions then this exam you would feel very simple. Because HadoopExam added few of the hard/complex exercises as well in all these practice questions.

If you have not practiced well then, this certification exam you would feel very difficult and you would not be able to complete the exercise in 2 hrs. If you have practice well then, most likely you would be able to complete real exam within 1 hr 30 mins.

Many of the learners are coming to HadoopExam and told us they have good knowledge of the Spark and few tasks they completed before the real exam, but they are not able to complete the exam on time. Which proves that practice before the real exam is necessary and this is again universal truth for all the exams.

Is it require to write complete application during real exam?

Many of you know that if you are writing Spark application using Scala, then you should have bundle that application using Scala Build Tool (SBT) or using Apache Maven. But this is not expected from you during the exam. Because this certification exam is not for build tool but rather testing your programming knowledge on the Spark framework. Hence, for both Spark-Scala and pyspark Hortonworks would give you already bundled or completely created application as below.

- **Spark-Scala**: A jar file would be bundled with all the required dependency and given to you.
- **Pyspark**: You would be getting a .py file with the complete application.

With this application you would be asked to submit the application on the yarn cluster with the spark-submit command and also required to use or later the existing configuration with the parameters.

Size of the data

Not all the tasks you would be given with the huge data. But rather smaller dataset would be given. However, out of all the tasks couple of tasks would involve huge data. And that may become challenging and time consuming as well. Data may contain 100's of parameters or columns in a csv file. You need to remove all the unwanted columns and apply join, filter and saving your final result. It is always recommended that all the easy questions should be attempted first and then go for high volume data. Because the cluster given to you most likely single node and not good enough for huge volume of the data.

Performance of the environment

Cluster provided in the cloud may not be performant but good enough for solving the given tasks. Hence, we have to be very careful when you submit any tasks and your solution must not involve the shuffle phase with huge volume of data. Because most of our learners have given the feedback that the cluster provided during the exam is very slow. However, since then it is improved and recently candidates are not facing this issue. There are occurrences where learners face the session disconnected issues during the exam, you may also be ready for such issues and it can be because of

- Your internet connection is not good
- Proctor internet connection is not good
- Cloud environment may not be reachable

Once your session got disconnected and connected back then you need to inform the proctor and he/she may deduct this time from your overall time. It all depend on the proctor discretion. If you belong to a country where internet connection reliability is challenging then make sure during your exam it does not happen.

Online/Offline Documentation provided during the exam

Hortonworks would provide some offline documentation which you can access during your exam. However, we highly recommend that you should not depend upon that documentation because rather than help you could spend more time searching things in the documentation and lose valuable time. The best thing is you remember few of the important things which you need frequently in your real exam for example

- Name of the class
- Name of the package
- Name of the functions and their arguments
- Creating and using case classes
- Creating custom types using Spark SQL (Practice this as much as possible) and remember all the types packages.

Should I know Pig, Sqoop, flume for HDPSCD Spark exam

No, you don't have to know this stuff at least for passing this certification exam. And there won't be any questions based on these topics. As we previously mentioned you should be comfortable in Spark, Hive, HDFS and command line.

Before answering any question, you must understand what is expected for completing the task. Otherwise we have seen that learners implement something else and then realize that they have implemented solution which was not asked and they again re-write the new solution and again and that is complete waste of time. Generally, the question is simple but, in the exam, they try to make it complicated by providing steps in different order. For instance, if question need to be solved in A, B, C and D order then they might give you detail in D, B, A, C order. Hence, please don't start thinking solution immediately until and unless you understand full task requirement.

Even from the feedback we get to know that in the question it is unnecessary highlighted few of the text which is not relevant and again you may get confused with that. Read the question twice and thrice before you start writing your script that you would correct things later on. Some of the questions are really straight forward as well in the real exam.

Another confusing stuff in the real exam is column number. If your solution starts with index 0 then make sure column number given in the question and in your solution matches. It is possible in the task they have considered start indexing from 1 but you implement based on 0.

Avoid typo during the exam and make yourself comfortable working with the "vi" editor, because in the exam we don't think you would be provided any modern text editor or IDE (Integrated development environment). Also don't depend on the auto complete or tabbing feature. If you have a habit of using Alt+Tab in the Windows then this may not work on the cloud environment because you are already in the Google chrome browser. Most of the time you would be using mouse.

This all above things which we have mentioned is based on the feedback we are receiving from last few years and things may have changed and environment may be improved by the time you go for the exam.

Can I prepare HDPSCD-Spark in two weeks?

Yes, it is possible. You need to spend around 4-6 hrs daily on the training and you should solve and understand all 89 questions provided by HadoopExam.com then you are ready to take this exam. However, it all depend on you. How much you have grasped and understood the stuff from the preparation material. Many of our learners completed this exam in less than 30 days. So, you can also do the same.

What is the proctor during real exam?

During your real exam there is one person who take care your real exam environment preparation as well as keep an eye on you. However, keep in mind that proctor is not only for keeping an eye on you but he or she is there to help you. If you find any issue with the connectivity, accessing material, checking time and any other status about the exam environment. They all are well trained for all these stuffs and very helpful.

You can start the exam 15 mins earlier than scheduled start time and proctor would ask you to show the desk and your place with the 360-degree using the webcam installed on either laptop or desktop. During the exam hours your desktop remain in sharing mode. It is always recommended you start 15 mins earlier than your scheduled time. You should always have a bigger monitor for your exam and avoid very small laptop screen and recommended size is 1600X900.

Linux Knowledge

It is always recommended that you have some basic Linux Knowledge and experience with the vi editor. It is not required that you have in depth Linux operating system knowledge. Make yourself aware about following Linux commands

- cd : Change directory
- ls : List all the directory and files at the current path
- vi : This is good editor for editing files in Linux
- rm : Remove a particular file
- mkdir : create directory
- cat : Read all the content of the file
- tail: continuous new output appended to the file.

Apache Ambari

In the Hortonworks instructions it is underlined that environment is using the HDP-3 with the latest version of Ambari-2.7. No need to worry about this. You would not be asked to do anything with the Apache Ambari in this particular certification. All the required services in the cluster already up and running. Ambari is a cluster management software, which is initially developed by Hortonworks and then later on it is open sourced under Apache Software Foundation.

How would be my exam day?

On the day of the exam, you would have to show your govt photo Identification using the webcam if you are appearing from home.

- **Browser**: You should have Google browser installed, and you would be asked to install one of the chrome extensions.
- **Chat with proctor**: Proctor would chat with you. You need to show your photo ID and 360-degree view of your room or area. He/she wanted to know and make sure that you are not making any recording as well as right candidate is appearing the exam.
- Proctor would launch or start an exam for you.
- Your entire exam would run under the Google chrome browser. And most likely you would be given AWS EC2 instance access.
- All the assignment or tasks are written on a file inside a folder or on desktop, and you can start solving the problem when your time starts.
- During the exam if you face any issue, you can chat with the proctor to solve the problem.
- There is no time tracker available on the exam window (learners' feedback), and you have to ask proctor in between for the remaining time.

- With each task/assessment/assignment there would be one empty text file. You need to write all your code in that text file and save it for evaluators to check.
- We see gradually based on the feedback they are improving the exam environment as well and now you can copy and paste the data as well, previously that was not possible.

Chapter-4: HDPSCD - Spark Syllabus

There are in total 22 different sub-topics which are part of the syllabus and divided in 3 major section. We will be discussing each one so that you can understand what is expected from you in the real exam and after that we would be providing some Hands exercising from the HadoopExam certification simulator.

HDPSCD Syllabus Section-1: Core Spark

In this section more focus would be given on your understanding of the Core Spark framework. Mainly, how spark distribution framework works. You are able to wisely use the RDD API or not. How to read and write data on the HDFS using the Spark RDD API. What is different modes client, cluster and local.

With the cluster mode you should be able to use yarn (yet another resource negotiator) scheduler and able to submit the already built application.

However, it is not expected from you that you have deep knowledge about yarn scheduler. (You can get in depth learning about this from our On Demand Hadoop Professional Training where each topic is discussed in depth). Let's start discussing each individual topic under this section.

Topic-1: Write a Spark Core application using Python or Scala

In this you would be writing an end to end Spark application using Scala or Python. Building end to end application is relatively easier in Python but not in the Scala. With the Scala you have to use

Scala build tool or maven to completely build an application and including all the dependencies (jars) in the same bundle.

However, from the feedback and experience we have seen that such tasks are not given in the real exam. Because this is the knowledge require for building end to end Scala application. And also, it is a quite huge tasks and cannot be easily achieved within 2 hrs, without using IDE. Hence, you would not be asked to bundle application using Maven or SBT. Rather you should be able to write complete Spark application which can be executed on the spark-shell or pyspark shell.

Topic-2: Initialize a Spark application

Under this subject you would be able to create SparkContext (In spark-shell or pyspark shell it is by default available with the variable name "sc") or SparkSession (available as "spark" variable in shell).

SparkContext require an object of SparkConf which has the information about your application. Also, make sure per JVM only one SparkContext object is created. Below is the sample Scala code to create and initialize the Spark application

```
val heConf = new SparkConf().setAppName("HadoopExam HDPSCD Spark").setMaster(master)

val heSparkContext = new SparkContext(heConf)
```

In pyspark it would be as below

```
heConf = SparkConf().setAppName("HadoopExam HDPSCD Spark").setMaster(master)

heSparkContext = SparkContext(conf=heConf)
```

Where

AppName: Name of your Spark Application, which would be printed in all your logs or on the Web GUI.

master: It could be anything from below

- **Local/spark**: provided by the Spark itself. Local should be used only for testing and unit testing.
- **Mesos**
- **Yarn**: For this certification this is the relevant one and you should always use this one for this certification. While using the spark-submit command also you have to use this one and **certainly you would get one of the tasks based on this**.

Topic-3: Run the Spark Job on YARN

Since Spark 0.6, YARN was supported and you would be able to run your Spark Jobs using this Yet another resource negotiator scheduler or cluster manager.

If you are doing cluster setup your own then you have to do some setting and update the properties to use the HDFS and YARN with the Spark. But for this certification preparation you would be using Sandbox either locally or in the cloud these all settings are already being done and you don't have to do it. You just need to know how to submit your application on the YARN with the spark-submit command line utility.

YARN is a cluster resource manager (RAM, CPU & DISK, currently only RAM is supported) and at the runtime resource allocated to your application based on the availability and demand (Learn more with the Hadoop Professional Training) . There are below two modes using which Spark application can be deployed.

- Cluster mode
- Client mode

In cluster mode, the Spark Driver runs inside an application master process, which is managed by YARN on the Hadoop cluster, and client do not have to be active once the application is launched and it can go away once the application is initiated.

With the client mode, driver runs in the client process itself and application master is only used for the requesting resources from the YARN.

Spark support many more cluster managers other than YARN as well. During the spark-submit command masters address would be provided using --master parameter. But for the YARN it requires the IP address of the ResourceManager or host name and that would be picked by the Spark from the Hadoop Configuration directly and you don't have to pass that IP address explicitly in the command line arguments. Hence, you just need to use "--master yarn". Below is the command for launching Spark application in the cluster mode on the YARN.

```
spark-submit --class org.hadoopexam.example.SparkPiApp
        --master yarn \
        -- deploy-mode cluster \
        -- driver-memory 4g \
        -- executor-memory 2g \
        -- executor-cores 1 \
        -- queues heQueue \
        Examples/jars/he-example.jar
```

In the above command it starts YARN client program and would be used for starting a default application master. There are some sample applications already provided with the Apache Spark Distribution which you can use to practice. In above command given application started as a separate thread as part of ApplicationMaster. In this case application client (Application Launcher) would periodically poll the application master for status updates and client would exit once the application finished running.

Similarly, if you want to launch application in client mode then start the spark-shell in the client mode with the yarn

spark-shell -- master yarn – deploy-mode client

Topic-4: Create an RDD

The fundamental data structure for the Spark is an RDD which is also known as Resilient Distributed Datasets, and that is fault-tolerant collection of elements that can be operated in parallel. There are basically two ways by which RDD can be created.

- Parallelizing collection in the driver program.
- Referencing the data stored in any external file system like
    - HDFS
    - HBase (NoSQL)
    - RDBMS
    - Hadoop Input Format

Parallelize the collection: In this case you have to use the SparkContext object and use the parallelize method with the existing collection in your driver program (For example Seq in Scala). This data or parallelized collection first partitioned and then copied over to the various nodes in the cluster. Below is the Scala example

val heData = Array ("Spark", "Scala", "Hadoop", "Python")

val heRDD = sc.parallelize(heData)

heRDD is an RDD, which is partitioned and distributed on various nodes on the cluster. And operations can be performed in parallel on each node in the cluster on the different part of the data.

Even you can define how many partitions you want to create while parallelizing the collection using the second arguments. Spark would run the code each partition in the cluster. Typically, best practice is that 2-4 partitions should be created for each CPU in your cluster.

Based on the cluster configurations Spark itself tries to set the number of partitions automatically. Below is the example for setting the 5 number of partitions explicitly.

sc.parallelize(heData,5)

Topic-5: Create an RDD from a file or Directory in HDFS

This is the main requirement where you can create RDD from external storage like HDFS, Cassandra, HBase, AWS S3, Azure blob etc. also, it supports the various efficient file formats for example text file, Sequence files, Parquet files, Avro files and various other Hadoop supported input format.

Even while creating or specifying external storage for the HDFS file system you can provide a single file for an entire directory, let's see one of the below examples for reading a single file

val heData = sc.textFile("HadoopExam.txt")

Important points

- If we are reading a file using local path then same path must exist on all the nodes in the cluster and file should be accessible on each node.
- You can provide single file; directory is supported compressed files as an input

sc.textFile("HadoopExam.txt") //loading single file
sc.textFile("he/hadoopexam") //loading all the contents for a directory
sc.textFile("HadoopExam.gz")//loading compressed data

with the second argument we can define number of partitions as well. By default, Spark creates one partition for each block of the file which is 128 MB or 256 MB in case of hdfs.

There are various other methods which can be used to read files for example

sc.wholeTextFile() : if directory contains more than one text file and you want to read filename as key and content as value. Then in this case using this function is a good idea. if we need to read single file then we are generally using *textFile()* method of the SparkContext object which return one record per line from the file. If you wanted to get to know the file name in your program then you can get the information from the RDD object. To read sequence file we need to use *sc.sequenceFile(K, V)* method.

However, in real exam

- You don't have to do any specific setup to read data from the hdfs rather you Just need to provide path of the file.
- You can also use new *SparkSession* object which was introduced in Spark 2.x, this is more convenient and has specific method to read various files and you can get *DataFrame* or *DataSet* object as a return value. And on this you can use relatively simpler API then RDD. HadoopExam.com has many exercises based on this, please go through them from this link.

Topic-6: Persist an RDD in memory or on Disk

We can persist or cache the RDD in the memory or on the disk. When RDD is persisted on each node locally, it stores the partition of the RDD and whenever computation initiated on the RDD then this would work on each partition on individual node. Benefit of Persisting and caching an RDD partition is that all the future operation on the RDD would be much faster. This is one of the basic requirements for machine learning and data science applications where iterative algorithms are used to run computation again and again on the same data. So, you would be caching the data which would be used again and again.

- To persist or cache the RDD we can use *persist* () or *cache* () methods. so, whenever first time action is called on RDD, that RDD would be persisted in memory or on the disk of each node. Remember each node would have different partition cached. RDD caching is fault tolerant, if we lose any of the partition because of node crash or node is not available. It is spark responsibility to recompute the rdd and get it back.

Spark supports various storage levels to store the RDD not only in memory cache but on the disk as well. You need to store intermediate RDD data on the disk. This can be achieved by providing storage level with the persist() method. If you use *rdd.cache()* method then by default it uses memory. Few of the storage types are below

- Memory-Only
- Memory-and-Disk
- Disk-only

However, this is not mentioned in syllabus, is there any specific storage level would be used or not. Most likely you would be asked to use the memory and *rdd.cache()* method which would suffice that requirement.

Spark automatically persist some intermediate data in shuffle operation for example during *reduceByKey*() method call.

Topic-7: Perform transformation on an RDD filtering & Aggregations.

Transformation create another RDD from the existing RDD. As you know RDD's are immutable. Hence, you have to apply transformation and that would return a new RDD. Some of the common functions you need to know which are transformation and would be used in your data pipeline

- map()
- filter()
- filterMap()
- union()
- intersection()
- distinct()
- groupByKey()
- reduceByKey()
- aggregateByKey()
- sortByKey()
- join()
- cogroup()
- pipe()
- coalesce()

In the syllabus it is talking about filtering and aggregation. However, this may not be limited in the exam and you may have to use above listed transformations as well. So better go and practice all the questions on the HadoopExam.com

RDD Filter:

Sometime data which are not relevant we need to filter out and you should do this as early as possible, because it can help you to save lot of processing time and disk space. Filtering is a transformation and give you a new filtered RDD.

Below are the few examples of the filter functions

```
val heData = sc.parallelize(List(0,1,2,0,5,9,10,0))

val  heFilterRDD = heData.filter(x => x!=0)
```

Here *heFiltereRDD* would have *[1,2,5,9,10]*

You can create your own custom functions as well to filter the data from RDD. If you have some complex logic then better approach is to create a separate function. Below are the few more examples of filter

```
heRDD.filter(x => x.getInt("key") >100)
```

heRDD.filter(line => line.contains("HadoopExam"))

Aggregations: Here aggregate does not mean to use only RDD aggregate functions/method but rather it is more a mathematical perspective like collecting all the data from RDD and compute summary e.g. sum(), max(), min(), count() etc. on numerical data.

Similarly, on the other data type you can apply summary like log aggregator, text aggregations etc.

As data is distributed across the machines, so final result needs to be aggregated from all the nodes/machines in the cluster. The main point is how efficiently data can be aggregated from all the distributed nodes in the cluster. Few examples of aggregations are below

val heGroup = heRDD.groupBy(x => x.value())

val count = heGroup.countByValue()

While aggregating data, it is required moving data over the network among all the machines in the cluster. Hence, always first filter the data which is not needed before applying the aggregations.

Topic-8: Perform Spark Actions on an RDD

To initiate the actual transformation which you have already written in a program require an action that is the reason transformation is called lazy. Example of the actions are below

heRDD.count()

heRDD.collect()

heRDD.saveAsTextFile()

Topic-9: Create and use broadcast variables and accumulators

Broadcast variable: Suppose your Spark program needs the data which is not part of the Spark RDD for example some lookup tables, small datasets etc and that needs to be joined with the RDD on each node. Then we can share this data on each node using the broadcast variable see the example below

Val heLookup = map(Hadoop->500, Spark->700, Scala ->500)

Val heRDD = sc.parallelize(Array (Hadoop, Spark, Scala))

In the above example *heLookup* is a shared variable which need to be sent over all the nodes in the cluster. To share that variable over all the nodes we need to broadcast it. Broadcast variable must be serializable, so that it can be sent to each node in the cluster. And executor on that node can use it locally. Broadcast variable can be created as below

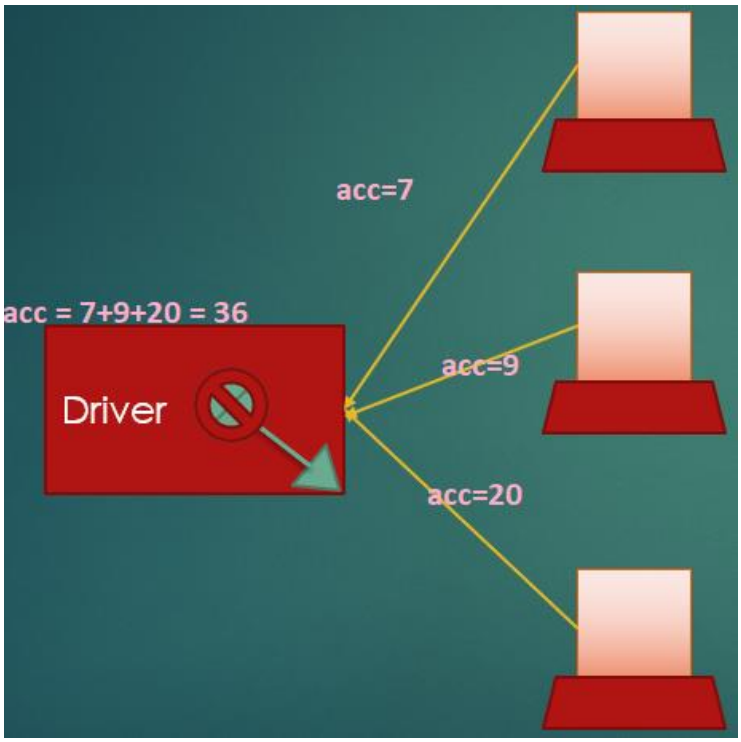
```
//Creating a broadcast variable
val heFeeDetail = sc.broadcast(map(Hadoop->500
                                , Spark->700
                                , Scala ->500)

//Now use the broadcast variable
val heRDD = sc.parallelize(Array(Hadoop, Spark, Scala)).join(heFeeDetail.value(_))
```

You can get the full-length exercise in HadoopExam online certification simulator.

If you want broadcast a variable, you have to use *value()* method on it. Learners sometime get confused why I am unable to broadcast an RDD. Because RDD is not broadcasted, if you have RDD then first convert it into Scala serializable data and then broadcast it, broadcast variable usually used for the read only purpose. If you change or update the values in the broadcast variable the same would not be replicated on another node. Broadcast variables are sent from driver to each node and this works one way only.

Accumulators: This is a shared variable which you can sent back to driver (which is not possible with the broadcast variable). This is updated on each node in the cluster and same update is propagated back to the driver (not on any other node in the cluster) and driver can accumulate all the individual nodes accumulator value and get the final concluded/resultant values. See the below block diagram to understand more.



Below is the sample code for using the accumulators

```
//you can create accumulator variable as below

val accCount = sc.accumulator(0)

//Use it on each node, with the same initial values

Val heResult = sc.parallelize(hugeCollection).map(x => {accCount + =1}).reduce((a,b) => a+b))
```

The final value of the accumulator can be received when you call *value()* method on it.

```
account.value()
```

Topic-10: Configure Spark properties

Before, after or during the Spark application deployment we should check that the required Spark properties are set or not. Even many properties we can set on Application level. There are various types of properties which we can set as below.

- Application level
  spark.app.name: Defining the name of the application which can be seen in the log and UI.

  spark.driver.memory: Driver process memory

- Runtime Environment property:
  spark.executor.extraJavaOptions: You can provide additional JVM properties e.g. GC settings.
- Shuffle behaviour: These all properties can be used to be applied some changes during the shuffle phase e.g.
  spark.shuffle.compress : Using this you can specify whether the compress the map output or not. If yes then you have to specify the codec as well using below property
  spark.io.compression.codec

  There are various such properties which can be specified, hence please know how to set all these properties because we cannot remember all the properties and for that we may have to go through the documentation as well. Below are the different sections for which properties can be defined.
  - Compression and Serialization
  - Memory management
  - Execution behaviour
  - Networking
  - Scheduling
  - Dynamic allocation
  - Security etc.

However, you would be given a particular property detail and you need to find the valid property name and use the same. You can set the Spark properties at below location

- **SparkConf**: Whatever property you set using the SparkConf it would be applied to an individual application parameter. Same properties can be configured using java system properties.
- **Environment variable**: These are machine/node level properties and can be set using conf/spark-env.sh file.
- **Logs level and log rollover settings**: Spark uses the Apache Spark log4J library and individual property can be set using ***log4j.properties*** file.

Below are some examples for each one of above.

Using *SparkConf* to set properties on the application level

```
val conf = new SparkConf().setMaster("yarn").setAppName("HE Application")

val sc = new SparkContext(conf)
```

Similarly, dynamically loading Spark properties: This is good way if you want to avoid hardcoding and do the certain configuration in a *SparkConf*. Suppose you want to run your application with different master, you can do as below.

```
val sc = new SparkContext(new SparkConf())
```

You would be writing above line in your application and use the below application to submit the code.

```
spark-submit -- name "HE App"
             -- master yarn
             -- conf spark.eventLog.enable=false
             -- conf "spark.executor.extraJavaOption=-   XX:+PrintGCDetails" HEApp.jar
```

In spark-shell and spark-submit below are the two ways in which properties can be set

- Command line option e.g.
  --master
- Using Configuration object
  --conf

There are some default properties as well, which are set using *conf/spark-default.conf* file. In this file properties can be specified using key and value.

Topic-11: Ingest data using SparkSession

SparkSession: This is an entry point for the Spark env. This is available since Spark 2.x version. Since then many things have changed for the Apache Spark. As we have seen before Spark 2.0, entry point was SparkContext

With the SparkContext our Spark application get access to the Spark cluster using the ResourceManager. As we have following resource managers available

- Local mode: Using number of threads as local[n]
- SparkStandAlone: Spark Default resource managers.
- Yarn : Hadoop Yet another resource negotiator.
- Mesos

Using SparkContext we can

- Cancel already submitted Job, hence it works as a handle for the submitted application.
- Set the configurations.
- Get the current status of the already submitted applications and few more things

We can still use the SparkContext object, mainly while working with RDD api and shared variables (Broadcast variable and Accumulators)

```
val conf = new SparkConf().serMaster(yarn).setAppName("HE App")
val sc = new SparkConf(conf)
```

SparkSession is a unification of various already available context like

- SQLContext
- Sparkcontext
- HiveContext
- StreamingContext

We can create SparkSession as below.

```
val SparkSession = SparkSession.builder()
                    .master("yarn")
                    .appName("HeApp")
                    .config("key", "value")
                    .getOrCreate()
```

We can use SparkSession to read the data.

```
val df = SparkSession.read.json("heData.json")
```

There is various format specific method available on SparkSession, which you should have practiced well before your real exam. Below are some of the examples

- *spark.csv()*
- *spark.jdbc()*
- *spark.json()*
- *spark.orc()*
- *spark.parquet()*
- *spark.text()*
- *spark.textFile()*

All above are the format specific methods. If you want to use format agnostic method then use load() method.

HadoopExam.com has SparkSQL Cookbook available to learn and understand SparkSQL in depth with 35 Hands On exercises)

Once the data is read then it would return DataFrameReader object.

Writing Data: Using SparkSession object we get the DataFrameReader object. But writer object we can get from DataFrame itself. You can use the code as below

heDF.write() : It would give us DataFrameWriter object.

Using the DataFrameWriter object, we can write data in various format on the local or HDFS or any other supported file system.

```
heDF.write.parquet("HEParquet")
heDF.write.csv("HECSV")
```

Topic-12: Sort results and write out to HDFS or other supported destinations.

It depends on what API you are using to save the data.

- SparkContext
- SparkSession

Both API provides different mechanism and methods to save the data on HDFS. Also, it depends on what API you are using from below

- RDD API
- SparkSQL API

We recommend that, until and unless specified you use SparkSQL API. This is easier as well as SaprkSQL is the future and many operations from the SparkSQL API are developer friendly then RDD API. Let's check few of the example below

```
val heRDD = sc.parallelize(Seq(
                    ("MySQL", 5500),
                    ("MLib", 5600),
                    ("ETL", 5700),
                    ("English",5800),
                    ("Spark1.0" , 5900),
                    ("Spark2.0", 6000)
                    ))
//Default sorting in ascending order
Val heSorted = heRDD.sortByKey()

//If you want to sort in descending order then use code below
Val heSortedRDD=heRDD.sortByKey(false)
```

Similarly, you can sort the data using the DataFrame as well. This is an example from Spark 2.0 version and going forward you should prefer this approach only.

```
val heLines = sc.textFile("heData.txt").toDF("line")
val df = heLines.explode("line", "word")(line.split(""))
val heSortedDF = df.groupBy("word").count().sort(desc("count"))
val heData = heSortedDF.show()
```

DataFrame API is much simpler than RDD API, as data is properly organized in column format, and can be easily sorted by just mentioning the name of the column on which sorting needs to be done.

Saving Data to HDFS

```
Val heDataDF = spark.read.load("hedata.parquet")
heDataDf.select("name","fee").write.save("heData.parquet")
```

We can save data in various format and use the option to specify the type of the data.

HDPSCD Syllabus Section-2: Spark SQL

Topic-13: Create Spark DataFrames from an existing RDD

Code written using the previous version of Spark (before Spark 2.0) mostly would have RDD API user. And you should be able to migrate that code on the Spark 2.0 framework. So sometime it may require you convert RDD into DataFrame. Below are some popular ways to convert RDD into DataFrame. DataFrame object uses Row object as first level of observation to represent data. Let's first create RDD and then create DataFrame from that.

```
val heRDD = sc.parallelize(
                    Seq(
```

```
("Hadoop", Array(1,2,3,4,5),
("Spark", Array(11,12,13,14,15)
("java", Array(21,22,23,24,25)
))
```

Creating DataFrame without schema

```
val df1 = spark.createDataFrame(heRDD)
df1.show()
```

Creating DataFrame by providing explicit schema and column name.

```
val df2= spark.createDataFrame(heRDD).toDF("CourseName", "Scores")
df2.show()
```

Explicitly specifying schema name

```
//Define a Case class for HadoopExam course detail
case class HECourse(id: Int, name: String, fee : Int, venue: String, duration:Int)

//Create an RDD with 5 HECourses
val courseRDD = sc.parallelize(Seq(
   HECourse(1, "Hadoop", 6000, "Mumbai", 5)
 ,HECourse(2, "Spark",5000, "Pune", 4)
 ,HECourse(3, "Python", 4000, "Hyderabad", 3)
 ,HECourse(4, "Scala", 4000, "Kolkata", 3)
 ,HECourse(5, "HBase", 7000, "Banglore", 7))
 )

//Check the types of RDD courseRDD
//Convert RDD into dataset, as RDD has schema information, so Dataset will automatically infer that schema.
val heCourseDS = courseRDD.toDS

//Type-1 Using Scala function (Lambda)
heCourseDS.filter(record => record.fee > 5000).show()

//Type-2 Column based SQL expression
heCourseDS.filter('fee > 5000).show()

//Type-3 As SQL DSL
heCourseDS.filter("fee > 5000").show()
```

Resilient Distributed Dataset

RDD is the lowest representation of data in Spark. Every processing in Spark done using RDD, whether you use SparkSQL abstractions like DataFrame/Dataset. An RDD spread across multiple machines in a Spark cluster, it provides APIs so you can work on it. You can create an RDD from different types of data source, e.g. text files, a database via JDBC, etc.

Apache Definitions of RDD: RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators.

DataFrame:

Similar to RDD, it is also distributed and immutable collections of data. You can imagine DataFrame as an RDBMS table with column name and rows. But DataFrame rows are divided and saved across various machines in Spark cluster as shown in below image.

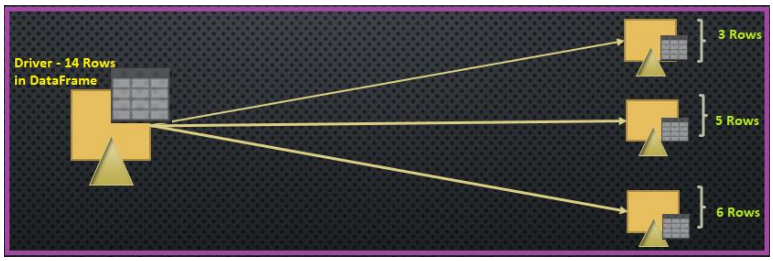


Figure 1: Partitioned DataFrame object across cluster nodes

- DataFrame helps in writing SparkSQL code using simpler API, and it is very similar to Python and R DataFrame.
- DataFrame is higher level abstraction of RDD.
- DataFrame represents Dataset with the generic Row object. So you can have below similarities between Dataset and DataFrame.

  ```
  DataFrame == Dataset<Row>
  ```

  Here Row is a generic object, and does not have type information attached to it.

Whenever you work with Dataset or DataFrame you are working with the Row objects. In case of DataFrame it can be generic Row object and in case of Dataset it will be typed Dataset objects.

Even, you can apply schema information to DataFrame object as well. To work with DataFrame you have following two approaches.

- SQL queries
- Query DSL (It can check the syntax at compile time)

Topic-14: Perform operations on the DataFrame

Topic-15: Write a Spark SQL application use Hive with ORC from Spark SQL

Topic-16: Write a Spark SQL application that reads and writes data from Hive tables.

Hive is older than Spark and previous project which was created in Spark to support Hive was known as Shark. However, Shark was only supporting Hive framework. So it was decided to not continue with the Shark and create a new project to support general purpose SQLs. And new project is Spark SQL, which is based on Catalyst optimizer.

Apache Spark and Hadoop framework complements each other on various front for example Hadoop provides distributed storage using HDFS and a matured cluster manager using YARN. And Spark is purely a computational framework which uses this distributed framework during computation and no new infrastructure needs to be created.

And most of the structured data was stored using Hive, and Hive has concept of external and managed tables. And all the information regarding these tables are stored in Hive catalog which is usually stored in another RDBMS solution like MySQL, Oracle etc.

- **Managed Tables or internal tables**:
  All the tables whose lifecycle is managed by Hive framework is known as internal or managed tables. Hive has one directory called or referred as warehouse directory. It is not necessary to have directory name to warehouse, but usually people keep the names as warehouse only for example /home/cloudera/warehouse, you will find on Cloudera platform. Hence, whatever tables are managed by Hive it will create a directory with the table name in this warehouse directory. For example, we have created a table called "TBL_HADOOPEXAM" then hive will create a directory as "/home/cloudera/warehouse/tbl_hadoopexam" and all the data will be stored in this directory. It may be stored in further nested directory, if table was created using partitions and buckets. As I mentioned internal tables are managed tables and managed by Hive. If you run the drop table DDL than table will be dropped as well as data and directory from HDFS "/home/cloudera/warehouse/tbl_hadoopexam" will be deleted.
  Internal tables have the following features.
  - Entire lifecycle of the Hive internal tables are managed by Hive only
  - As soon as you drop table it will drop the data as well as metadata associated with the tables from both the metastore and namenode of HDFS.
  - Data will always be stored under warehouse directory.
  - You can have any directory name as a warehouse directory, by setting the configuration on Hadoop using "**hive.metastore.warehouse.dir**"
  - Even you can create managed table other than warehouse directory, while creating table you can provide the location where you want to create the table. But if you don't want to specify the location than Hive managed table data will always be created in warehouse directory.

- **External Tables**:
  In case of external tables data lifecycle is not managed by Hive and data should not be stored in warehouse directory. You can have your data stored in HDFS in any other directory. Even existing data you can convert into external tables without moving them into warehouse directory. When you define an external table than only metadata for the table will be created in the metastore. Following are the features of Hive external table

  - Any existing data in HDFS can be created as a Hive table without moving them into warehouse directory.
  - If you drop Hive external table than data will not be deleted and remain as it is on HDFS.
  - If you drop table than metadata of the table will be removed from metastore.

Hive Metastore:

While creating the tables you have to provide that what is the location of the data, what is name of table, what is the name of columns, types of columns, data partitioning, bucketing, ser-de (serializes and de-serializes) detail. These all information needs to be stored somewhere and that will be stored

in Hive metastore. You can use RDBMS like Oracle or MySQL to store this metadata. If you don't configure than Hive will be using in-process instance of Derby (which is not good for production use cases).

Hive Support in SparkSQL

Developer of SparkSQL wanted to still continue to support HiveQL. Hence it was maintained and SparkSQL supports features of Hive like reading Hive table's metadata from metastore, Hive Query Language and Hive User defined functions etc. To have Hive support with SparkSQL, you don't have to do any changes in your Hive data and Hive metastore.

Some of the features which are not supported from Hive are below (You can find full list from this link):

- Hive tables bucketed using Hash Partitioning
- Hadoop Archival
- Hive optimization using Index. (Because SparkSQL stores data in-memory while doing computation. Hence, there is no as such need of doing indexing on the data stored in Hive)
- **Merging Small files**: If Spark SQL query generates lot of small files than it will not merge all the small files. But Hive has this optimization features. Because having lot of small files are not good for HDFS.
- Not all the UDF (User defined functions) from Hive are supported.

Hive Query support using SparkSQL:

SparkSQL supports various Hive features as well as HiveQL, and please note that it is not necessary to have Hive setup already available to run Hive Query or to use Hive features. You can use all the supported Hive features and HiveQL without even setting up Hive. Hence, this will give you the advantage if you have already worked with Hive and familiar with Hive features and wants to use them in SparkSQL.

SaprkSQL supports the queries written using HiveQL. HiveQL is very similar to SQL, but does not follow ANSI standard of SQL. Hence, you can say that HiveQL is a SQL-like language. The reason HiveQL is supported because HiveQL is quite old and more matured than SparkSQL, even it has support for more complex queries than SparkSQL. We will see one example, how to use HiveQL in SparkSQL in next section. It internally uses the HiveContext and launches the Spark Jobs to run HiveQL.

In the below example the syntax which we are seeing "FROM HEVIEW" is a Hive way of writing SQL and that is supported without even setting the Hive.

Load Data from file, this also shows that SparkSQL supports Hive Query Language Syntax

```
//Load data from a file, and as file has header. So by providing options you can say, derive column from file header only.
sql("CREATE OR REPLACE TEMPORARY VIEW HEVIEW USING csv OPTIONS ('path'='/home/hadoopexam/spark2/sparksql/HadooExam_Training.csv', 'header'='true')")

//Select all the data from View, using HiveQL syntax
sql("FROM HEVIEW").show

//Print schema detail for HEVIEW
```

```
sql("desc EXTENDED HEVIEW").show()
```

```
scala> sql("CREATE OR REPLACE TEMPORARY VIEW HEVIEW USING csv OPTIONS ('path'='/home/hadoopexam/spark2/sparksql/HadoopExam_Training.csv', 'header'='true')")
2018-08-29 22:29:42 WARN  ObjectStore:568 - Failed to get database global_temp, returning NoSuchObjectException
res0: org.apache.spark.sql.DataFrame = []

scala> sql("FROM HEVIEW").show
+---+--------------------+----+---------+-----------+--------+
| ID|                Name| Fee|    Venue|       Date|Duration|
+---+--------------------+----+---------+-----------+--------+
|  1|          HE Hadoop| 9000|   Mumbai|01-Aug-2018|       2|
scala> sql("desc EXTENDED HEVIEW").show()
+--------+---------+-------+
|col_name|data_type|comment|
+--------+---------+-------+
|      ID|   string|   null|
|    Name|   string|   null|
|     Fee|   string|   null|
|   Venue|   string|   null|
|    Date|   string|   null|
|Duration|   string|   null|
+--------+---------+-------+
```

Topic-17: invoke SQL API or SparkSession SQL functionality to select and produce results.

Topic-18: Using Join capabilities to produce analytic results.

Joins Introduction: A Join is a way to retrieve information from two or more datasets. There are various types of joins. A normal JOIN, which is also called an INNER JOIN, a LEFT OUTER JOIN, a RIGHT OUTER JOIN, a FULL OUTER JOIN and CROSS JOIN.

SQL Example of inner join

Suppose a you wanted to know what employee worked in what department. While someone could just compare the ID numbers between the two tables, a way to have the information in one place is by doing a JOIN, also known as an INNER JOIN. Because they have one type of data in common, the department ID, the tables can be joined together.

```
SELECT LastName, DepartmentName FROM employee join department on
department.DepartmentID = employee.DepartmentID;
```

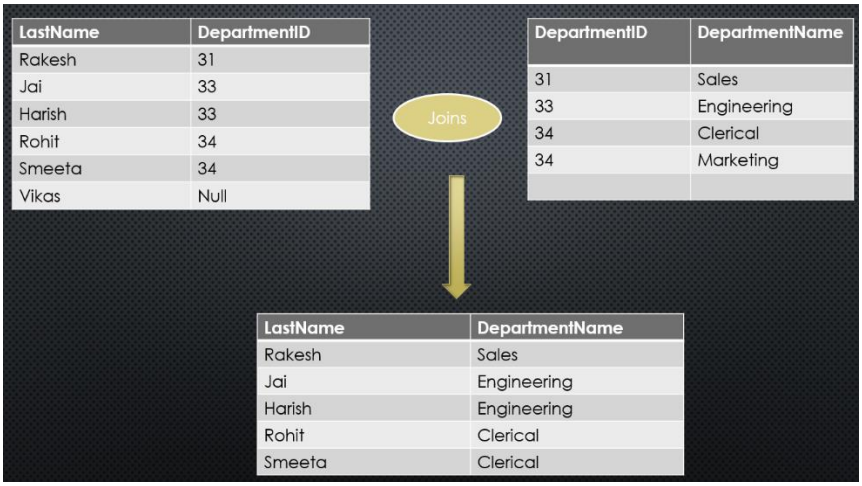


Figure 2: SQL Inner Join example

Outer Joins: Inner joins are fine if both tables have a matching record. However, if one table does not have a record for what the join is being built on, the query will fail. But if a database programmer

needs to grab information in an event that there is not a matching record for a row on one of the tables, they need to use an outer join. Types of outer joins are

- Left
- Right
- Full outer join
- Cross Joins

We will be doing joins example using SparkSQL. However, concept for joining dataset in SparkSQL and tables in SQL Databases are same. So if you have ever done this things in RDBMS then it would be quite easy for you.

Explanation for

- **Left Join**: A left outer join (also known as a left join) will contain all records from the left dataset, even if the right dataset does not have a matching record for each row.
- **Right Join**: A right outer join works almost like a left outer join, except with how the datasets are handled reversed. This time, all of the relevant information will be returned from the right dataset, even if the left table does not have a matching result. If the left dataset does not have a matching result, null will be in the place of the missing data.
- **Full outer join**: The FULL OUTER JOIN return all records when there is a match in either left dataset or right dataset records.
- **Cross Join**: The CROSS JOIN produces a result set which is the number of rows in the first dataset multiplied by the number of rows in the second dataset if no WHERE clause is used along with CROSS JOIN. This kind of result is called as Cartesian Product.
- Spark Joins Hands on Exercises:

Exercise: Spark SQL Dataset Joins

```
//Lets create two datasets
val heDF1 = spark.read.format("csv").option("header",true).option( "Inferschema",
true).load("/home/hadoopexam/spark2/sparksql/HadooExam_Training.csv")

//Create another Dataset
val heDF2= sc.parallelize(Seq( (1, "Hadoop", 6000, "Mumbai", 5), (2, "Spark", 5000, "Pune", 4), (3,
"Python", 4000, "Hyderabad", 3))).toDF("ID","Name","Fee","City","Days")

//Inner Join
heDF1.join(heDF2, "ID").show()

//Left Join
heDF1.join(heDF2, Seq("ID") , "left").show()
```

```
scala> val heDF1 = spark.read.format("csv").option("header",true).option( "Inferschema", true).load("/home/hadoopexam/spark2/sparksql/HadooExam_Training.csv")
2018-08-31 23:42:41 WARN  ObjectStore:568 - Failed to get database global_temp, returning NoSuchObjectException
heDF1: org.apache.spark.sql.DataFrame = [ID: int, Name: string ... 4 more fields]

scala> val heDF2= sc.parallelize(Seq( (1, "Hadoop", 6000, "Mumbai", 5), (2, "Spark", 5000, "Pune", 4), (3, "Python", 4000, "Hyderabad", 3))).toDF("ID","Name","Fee","C
y","Days")
heDF2: org.apache.spark.sql.DataFrame = [ID: int, Name: string ... 3 more fields]

scala> heDF1.join(heDF2, "ID").show()
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
| ID|       Name| Fee|    Venue|       Date|Duration|  Name| Fee|     City|Days|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
|  1|  HE Hadoop|9000|   Mumbai|01-Aug-2018|       2|Hadoop|6000|   Mumbai|   5|
|  2|   HE Spark|7000|  Kolkata|04-Aug-2018|       3| Spark|5000|     Pune|   4|
|  3|HE SparkSQL|6000|Hyderabad|07-Aug-2018|       4|Python|4000|Hyderabad|   3|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+


scala> heDF1.join(heDF2, Seq("ID") , "left").show()
+---+--------------------+-----+----------+-----------+--------+------+----+------+----+
| ID|                Name|  Fee|     Venue|       Date|Duration|  Name| Fee|  City|Days|
+---+--------------------+-----+----------+-----------+--------+------+----+------+----+
| 31|            HE LDAP| 7000| Singapore|31-Aug-2018|       4|  null|null|  null|null|
| 85|           HE DotNet| 7000|    Mumbai|13-Aug-2018|       2|  null|null|  null|null|
```

//Right Join

heDF1.join(heDF2, Seq("ID") , "right").show()

//Full outer Join

heDF1.join(heDF2, Seq("ID") , "fullouter").show()

```
scala> heDF1.join(heDF2, Seq("ID") , "right").show()
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
| ID|       Name| Fee|    Venue|       Date|Duration|  Name| Fee|     City|Days|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
|  1|  HE Hadoop|9000|   Mumbai|01-Aug-2018|       2|Hadoop|6000|   Mumbai|   5|
|  2|   HE Spark|7000|  Kolkata|04-Aug-2018|       3| Spark|5000|     Pune|   4|
|  3|HE SparkSQL|6000|Hyderabad|07-Aug-2018|       4|Python|4000|Hyderabad|   3|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+


scala> heDF1.join(heDF2, Seq("ID") , "fullouter").show()
+---+--------------------+-----+----------+-----------+--------+------+----+------+----+
| ID|                Name|  Fee|     Venue|       Date|Duration|  Name| Fee|  City|Days|
+---+--------------------+-----+----------+-----------+--------+------+----+------+----+
| 31|            HE LDAP| 7000| Singapore|31-Aug-2018|       4|  null|null|  null|null|
| 85|           HE DotNet| 7000|    Mumbai|13-Aug-2018|       2|  null|null|  null|null|
```

//Broadcast Join using function

heDF1.join(broadcast(heDF2), "ID").show()

```
scala> heDF1.join(broadcast(heDF2), "ID").show()
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
| ID|       Name| Fee|    Venue|       Date|Duration|  Name| Fee|     City|Days|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
|  1|  HE Hadoop|9000|   Mumbai|01-Aug-2018|       2|Hadoop|6000|   Mumbai|   5|
|  2|   HE Spark|7000|  Kolkata|04-Aug-2018|       3| Spark|5000|     Pune|   4|
|  3|HE SparkSQL|6000|Hyderabad|07-Aug-2018|       4|Python|4000|Hyderabad|   3|
+---+-----------+----+---------+-----------+--------+------+----+---------+----+
```

//Define a Case class for HadoopExam course detail
//Using JoinsWith operator

```
case class HECourse(id: Int, name: String, fee : Int, venue: String, duration:Int)

val heDS1 = sc.parallelize(Seq(HECourse(1, "Hadoop", 6000, "Mumbai", 5),HECourse(2, "Spark",
5000, "Pune", 4),HECourse(3, "Python", 4000, "Hyderabad", 3) ,HECourse(4, "Scala", 4000,
"Kolkata", 3),HECourse(5, "HBase", 7000, "Banglore", 7) ,HECourse(4, "Scala", 4000, "Kolkata",
3),HECourse(5, "HBase", 7000, "Banglore", 7) ,HECourse(11, "Scala", 4000, "Kolkata",
3),HECourse(12, "HBase", 7000, "Banglore", 7))).toDS()

val heDS2 = sc.parallelize(Seq(HECourse(1, "Hadoop", 6000, "Mumbai", 5),HECourse(2, "Spark",
5000, "Pune", 4),HECourse(3, "Python", 4000, "Hyderabad", 3))).toDS()
```

//Now apply the joinsWith operation, it will help you to provide the required conditions

//apply inner join

```
val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID"))
resultDS.show
resultDS.printSchema
```

```
scala> case class HECourse(id: Int, name: String, fee : Int, venue: String, duration:Int)
defined class HECourse

scala>

scala> val heDS1 = sc.parallelize(Seq(HECourse(1, "Hadoop", 6000, "Mumbai", 5),HECourse(2, "Spark", 5000, "Pune", 4),HECourse(3, "Python", 4000, "Hyderabad", 3) ,HECou
se(4, "Scala", 4000, "Kolkata", 3),HECourse(5, "HBase", 7000, "Banglore", 7) ,HECourse(4, "Scala", 4000, "Kolkata", 3),HECourse(5, "HBase", 7000, "Banglore", 7) ,HECou
se(11, "Scala", 4000, "Kolkata", 3),HECourse(12, "HBase", 7000, "Banglore", 7))).toDS()
heDS1: org.apache.spark.sql.Dataset[HECourse] = [id: int, name: string ... 3 more fields]

scala>

scala> val heDS2 = sc.parallelize(Seq(HECourse(1, "Hadoop", 6000, "Mumbai", 5),HECourse(2, "Spark", 5000, "Pune", 4),HECourse(3, "Python", 4000, "Hyderabad", 3))).toDS
()
heDS2: org.apache.spark.sql.Dataset[HECourse] = [id: int, name: string ... 3 more fields]

scala>

scala> val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID"))
resultDS: org.apache.spark.sql.Dataset[(HECourse, HECourse)] = [_1: struct<id: int, name: string ... 3 more fields>, _2: struct<id: int, name: string ... 3 more fields
]

scala> resultDS.show
+--------------------+--------------------+
|                  _1|                  _2|
+--------------------+--------------------+
|[1, Hadoop, 6000,...|[1, Hadoop, 6000,...|
|[3, Python, 4000,...|[3, Python, 4000,...|
|[2, Spark, 5000, ...|[2, Spark, 5000, ...|
+--------------------+--------------------+
```

//apply Left join

```
val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID") , "left")
resultDS.show
resultDS.printSchema
```

//apply right join

```
val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID") , "right")
resultDS.show
resultDS.printSchema
```

```
scala> val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID") , "left")
resultDS: org.apache.spark.sql.Dataset[(HECourse, HECourse)] = [_1: struct<id: int, name: string ... 3 more fields>, _2: struct<id: int, name: string ... 3 more field
]

scala> resultDS.show
+--------------------+--------------------+
|                  _1|                  _2|
+--------------------+--------------------+
|[12, HBase, 7000,...|                null|
|[1, Hadoop, 6000,...|[1, Hadoop, 6000,...|
|[3, Python, 4000,...|[3, Python, 4000,...|
|[5, HBase, 7000, ...|                null|
|[5, HBase, 7000, ...|                null|
|[4, Scala, 4000, ...|                null|
|[4, Scala, 4000, ...|                null|
|[11, Scala, 4000,...|                null|
|[2, Spark, 5000, ...|[2, Spark, 5000, ...|
+--------------------+--------------------+


scala> val resultDS = heDS1.joinWith(heDS2 , heDS1("ID") === heDS2("ID") , "right")
resultDS: org.apache.spark.sql.Dataset[(HECourse, HECourse)] = [_1: struct<id: int, name: string ... 3 more fields>, _2: struct<id: int, name: string ... 3 more field
]

scala> resultDS.show
+--------------------+--------------------+
|                  _1|                  _2|
+--------------------+--------------------+
|[1, Hadoop, 6000,...|[1, Hadoop, 6000,...|
|[3, Python, 4000,...|[3, Python, 4000,...|
|[2, Spark, 5000, ...|[2, Spark, 5000, ...|
+--------------------+--------------------+
```

Broadcast Join

In this join one of the datasets will be broadcasted to the nodes, which are holding partition of another bigger dataset. This type of join is known as replicated join because smaller dataset will be replicated on the other node in Spark Cluster.

Smaller dataset will be broadcasted. How do you find that the dataset is smaller one and needs and can be broadcasted? There is a configuration parameter which is set with the default value of 10MB. If dataset is less than 10 MB in size than it will be broadcasted. Parameter is as below.

spark.sql.autoBroadCastJoinThreshold

In broadcast join always smaller dataset should be broadcasted, so that network IO will be lesser. If you send large data over the n/w then it would be a bigger overhead. Sometime it is also known as Map-side join mostly in Hadoop world. Because joins is accomplished using just Mapper part of the Map-Reduce framework.

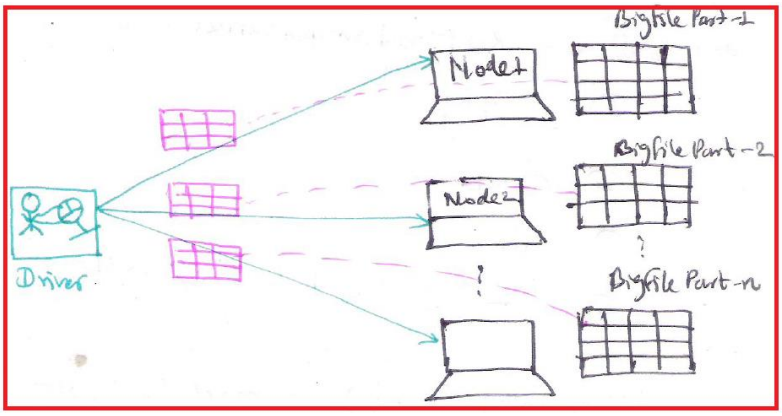


Figure 3: Broadcast Join

Driver is responsible to sending the smaller dataset over the network to each node on the cluster where bigger chunk of the data is residing and join will be locally applied on that node.

Topic-19: Rename DataFrame/Dataset columns to produce best results.

HDPSCD Syllabus Section-3: Spark Streaming

Topic-20: Use Spark structured streaming to ingest data in real time

Structured Streaming: You must understand the basic concepts of structured streaming and how some complicated things of streaming data can be easily achieved using the structured streaming API. Structured streaming solution was developed from scratch to use SparkSQL features and is recommended as well you should not use the old DStream API for the structured data which was part of Spark 1.x

One of the Major things about the structured streaming is that you can express your streaming computation the same way as you would express batch computation on static data.

Same SparkSQL engine would be Used to run streaming comp computation incrementally and continuously. we can use the same API of DataSet or Dataframe to apply transformation and aggregation on streaming data.

internal structure streaming is also executed using Spark SQL engine which is converted as micro batch processing. this micro batches can give 100 millisecond latency. there are various sources supported for ingesting data in the stream as mentioned below

1. **File source**: read and Fetch the files as soon as they are written to a directory and process that data. currently supported file formats at text, csv, JSON, ORC, Parquet etc. DataStreamReader interface would be used to read the data. the only requirement is that file written or placed in a directory must be done atomically.
2. **Kafka Source**: read data from Kafka messaging engine
3. **Socket Source**: it supports reading UTF 8 text format from socket connection. however, it is suggested that you use this more only for testing, because using this source it is not possible to provide end-to-end fault tolerance guarantee.
4. **Rate Source**: this is again used for testing purpose to generate specified number of rows per second, each output row contains a timestamp and a value. if you want to do benchmarking or testing then this is one of the good options to use.

Each type of input source comes with different options which you should be aware how to use and when to use.

Below is an example of reading data from the socket source

```
//First you need to start a Data server (netcat)
 nc -lk 9999

//connect to the Stream, which is generated at the socket on port 9999
val lines = spark.readStream.format("socket").option("host", "localhost").option("port", 9999).load()
```

Below is an example of reading data from the file system

```
//Create a Stream which can read the data from a particular HDFS //directory. As we know our data are csv data, even they are given in text //file. So use specific method to read csv data.

val trainingStreamDF = spark.readStream.option("sep", ",").option("header",false).schema(heSchema).csv("stream2/")
```

Topic-21: Invoke Streaming transformations and aggregations to produce analytic results.

Topic-22: Invoke spark-submit utility on existing Spark application using proper arguments.

Chapter-5: Sample hands-on exercises for the HDPSCD Spark Scala

In this chapter we will be providing 10 hands on exercise for this certification including the data. Which you can execute on the Hortonworks Sandbox. If you follow the comments in each exercise then you can easily understand the solution. In this book we would be having Problem statement and then step by step solution for the problem. We would be covering almost 10 exercises in this book out of all 89 exercises.

Exercise-1: You have been given txt file named hadoopexam1.txt with the following content.

hadoopexam1.txt

```
I am learning Apache Spark from HadoopExam Learning Resources
I am learning Apache Hadoop from HadoopExam Learning Resources
I have created my technical profile at www.QuickTechie.com
I am learning Apache Spark from Training4exam Learning Resources
```

Now accomplish the following activities.

- Create this txt file in hdfs.
- Once file is created, write a Spark application which will read this file from hdfs as Dataset.
- Once Dataset loaded, do the line count as well as fetch the first record from the Dataset.

Solution:

Step 1: Create a local director named "hdpcd"

mkdir hdpcd

Step 2 : Create a file locally "hdpcd/hadoopexam1.txt" with the given content.

Step 3 : Now upload this file to HDFS under

"hdpcd/hadoopexam1.txt"
hdfs dfs -mkdir hdpcd
hdfs dfs -put hadoopexam1.txt hdpcd/

Step 4 : Now in spark shell load this file as an RDD.

//Command to start spark shell :
spark-shell

Step 5 : Write spark code using Scala as being asked.

val heData = spark.read.textFile("hdpcd/hadoopexam1.txt") //Load the text file using SparkSession object.

// Count the number of lines in the Dataset
heData.count()

// Get the first record from the Dataset
heData.first()

Exercise-2: You have been given a txt file named hadoopexam2.txt with the following content.

hadoopexam2.txt

I am learning Apache Spark from HadoopExam Learning Resources
I am learning Apache Hadoop from HadoopExam Learning Resources
I have created my technical profile at www.QuickTechie.com
I am learning Apache Spark from Training4exam Learning Resources
I am learning Apache Spark from Training4exam Learning Resources

Now accomplish the following activities.

- Create this txt file in hdfs.
- Once file is created, write a Spark application which will read this file from hdfs as an RDD.
- Filter all the records which contains 'HadoopExam' in line and count the lines.
- Filter all the records which does not contains 'HadoopExam' in line and count the lines.

Solution:

Step 1 : Create a file locally "hdpcd/hadoopexam2.txt" with the given content.

Step 2 : Now upload this file to HDFS under *"hdpcd/hadoopexam1.txt"*

hdfs dfs -put hadoopexam2.txt hdpcd/

Step 3 : Now in spark shell load this file as an RDD.

Command to start sprk shell :

spark-shell

Step 4 : Write spark code using Scala as being asked.

// Create an RDD called lines, make sure you have a correct //path to file.

val lines = sc.textFile("hdpcd/hadoopexam2.txt")

//Apply the filter transformations and get the lines which //contains the word HadoopExam

val hadoopexamLines = lines.filter(line => line.contains("HadoopExam"))

//Apply the count action to get the total number of rows, //after the filter.

hadoopexamLines.count()

//Now we need to other way round, find all the rows which //does not have //HadoopExam keyword in it.// You can use //! symbol to make result as expected.

val hadoopexamLinesNotMatch = lines.filter(line => !line.contains("HadoopExam"))

//Apply the count action to get the total number of rows, //after the filter. //As in our file there were 5 records, we //expect count as 3.

hadoopexamLinesNotMatch.count()

Exercise-3 You have been given following list of words.

"We", "Are" ,"Learning" , "Hadoop" , "From" , "HadoopExam" , "We", "Are" ,"Learning" , "Spark" , "From" , "HadoopExam.com" , "hadoop" , "HADOOP"

Please accomplish the following activities.

- Create an RDD using for these words.
- Once RDD is created count all the words.
- Now filter out all the words which does have Hadoop keyword, however make sure it would count all the different cases(upper/lower) as well.

Solution:

Step 1: Now create an RDD using the given words.

//You have to first create a List using all the data given. If it is already in file, then use the file read approach. //If file is not given then create List and then parallelize (Distributed Data on all the nodes in the cluster) the List.

val lines= sc.parallelize(List("We", "Are" ,"Learning" , "Hadoop" , "From" , "HadoopExam" , "We", "Are" ,"Learning" , "Spark" , "From" , "HadoopExam.com" , "hadoop" , "HADOOP"))

Step 2: Now count the words in RDD

//Once RDD is created, we can apply the count action on it. //To get the total number of words in RDD

lines.count()

Step 3: Now filter the words from RDD as requested using the filter transformation.

//We are converting all the words in lower case and then //filtering.

val hadoopexamLines = lines.filter(line => line.toLowerCase().contains("hadoop"))

hadoopexamLines.count()

Exercise-4: You have been given following content in three different files.

hadoopexam4A.txt

HadoopExam.com QuickTechie.com Training4Exam.com

hadoopexam4B.txt

Hadoop Spark Scala Python Java Cloud Science

hadoopexam4C.txt

India USA UK Canada Australia

Now do the following activities.

- Load all three files in three different RDDS
- Concatanate all the data in single RDD
- Count all the words in of all three files.

Solution:

Step 1: Create three files locally, with the given content.

hadoopexam4A.txt

HadoopExam.com QuickTechie.com Training4Exam.com

hadoopexam4B.txt

Hadoop Spark Scala Python Java Cloud Science

hadoopexam4C.txt

India USA UK Canada Australia

Step 2 : Now upload all these files in hdfs.

hdfs dfs -put hadoopexam4A.txt hdpcd/

hdfs dfs -put hadoopexam4B.txt hdpcd/

hdfs dfs -put hadoopexam4C.txt hdpcd/

Step 3: Now write spark code to load these files

// Create an RDD called lines, make sure you have a correct path to file.

```
val linesA = sc.textFile("hdpcd/hadoopexam4A.txt")

val linesB = sc.textFile("hdpcd/hadoopexam4B.txt")

val linesC = sc.textFile("hdpcd/hadoopexam4C.txt")
```

//Now split the lines based on the space. So that after //transformation, new RDD available containing words from //the line.

```
val linesA1 = linesA.map(line => line.split(" "))

val linesB1 = linesB.map(line => line.split(" "))

val linesC1 = linesC.map(line => line.split(" "))
```

Step 4: As it was asked to combine all the words across RDD. Now union all three RDDS

//There is some confusion whether to use union or unionAll.

//Spark 2.0 onwards, both are same. Both keep the duplicate values/

```
val linesAB = linesA1.union(linesB1)

val linesABC = linesAB.union(linesC1)
```

//If you want to check the content in between then use the collect action.

//Remember, if RDD is very Big then it can cause OOM (out of memory) issue.

//Because it collect all the data on the Driver (Check HadoopExam.com Spark training for more detail)

//This is the step, where the code till now, you have written would be called.

linesABC.collect()

Step 5: Now flatten the RDD

//As you can see in the above step, it contains embeded Array objects for each RDD.

//We wanted all of them to be adjusted in the single Array and no embedded Arrays.

//Hence, we would be using the flatMap transformation, which will flatten all the words in single RDD.

val flatData = linesABC.flatMap(y => y)

//Check the content of RDD

flatData.collect()

Step 6: Check the count of data

flatData.count()

Exercise-5: You have been given following data in two different directories as below.

hdpcd/dir5/hadoopexamA.txt

HadoopExam.com QuickTechie.com Training4Exam.com

hdpcd/dir5/hadoopexamB.txt

Hadoop Spark Scala Python Java Cloud Science

hdpcd/dir5C/hadoopexamC.txt

India USA UK Canada Australia

Now accomplish the following activities.

- Load the data given in all three directories.
- Once the data is loaded filter the words which contain "Hadoop"
- Now persist the data in memory.

Solution:

Step 1: Create files locally, with the given content.

Step 2: Load this file in hdfs.

Create directory in hdfs.

hdfs dfs -mkdir hdpcd/dir5

hdfs dfs -mkdir hdpcd/dir5C

Load the files

hdfs dfs -put hadoopexamA.txt hdpcd/dir5/

hdfs dfs -put hadoopexamB.txt hdpcd/dir5/

//Be cautious, we need to upload 3rd file in separate directory.

hdfs dfs -put hadoopexamC.txt hdpcd/dir5C/

Step 3: Now Write the spark code

// Create an RDD called lines, make sure you have a correct path to file.

//Here we are not pointing to directory explicitely. Rather loading entire contents from the directory.

val linesA = sc.textFile("hdpcd/dir5/")

//Split line with the space. So that words of RDD can be created.

val linesA1 = linesA.map(line => line.split(" "))

// Create an RDD called lines for data from other directory.

val linesC = sc.textFile("hdpcd/dir5C")

//Split line with the space. So that words of RDD can be created.

val linesB1 = linesC.map(line => line.split(" "))

Step 4: Now collect the data and verify the contents

linesA1.collect()

linesB1.collect()

Step 5: Now persists the RDD.

//There are various persist level. Default is MEMORY_ONLY.

//There are more, you can check in Spark Training by HadoopExam.com

linesA1.persist()

linesB1.persist()

linesA1.count()

linesB1.count()

//Other way to cache the RDD

linesA1.cache()

linesB1.cache()

//Until you call action, RDD would not be cached.

linesA1.count()

linesB1.count()

Exercise-6: You have been given a following data in a file hadoopexam6.txt.

CourseName,Price,TaxandOthersInPercent
Hadoop,2900,15
Spark,3500,14
AWS,2700,13
Azure,2800,11
Java,3000,16
HBase,3200,20

Accomplish the following activities.

- Load this csv file in RDD
- Now Calculate the final price, using tax
- Save the final data on hdfs.

Solution:

Step 1 : Create file locally hdpcd/hadoopexam6.txt.

//Make sure, it does not include the Header.

Step 2 : Copy this file to hdfs

hdfs dfs -put hadoopexam6.txt hdpcd/

Step 3: Now write the Spark code as asked to solve the problem.

//Load the csv files as RDD. We are not explicitly reading as csv.

//It's loaded as lines.

val linesA = sc.textFile("hdpcd/hadoopexam6.txt")

#Check data is correctly loaded or not

linesA.collect()

#Now calculate final price #First Split the values with "," #Convert the 3rd column value to double. Hence, we can #divide it by 100 as line.split(",")(2).toDouble/100 #Add the #calculated price to actual price #line.split(",")(1).toInt)+((line.split(",")(1).toInt)*(line.split(",")(2).toDouble/100)) #Now finally append the final price in same line. Entire solution will be in single line as below.

//If it is not explicitly asked to use RDD, we should avoid using RDD. And rather use the DataFrame.

//As this exam has quite a good amount of focus on RDD. So learn RDD action and transformation as well.

//If you plan to contribute the Spark opensource framework, then you must know the RDD action and transformation as well.

//We will do exercises with the DataFrame and Dataset in later sessions.

```
val linesA1 = linesA.map(line =>
(line,(line.split(",")(1).toInt)+((line.split(",")(1).toInt)*(line.split(",")(2).toDouble/100))))
```

//Check data is correctly loaded or not

```
linesA1.collect()
```

Step 4: Now save the data as a text file in hdfs.

```
linesA1.saveAsTextFile("hdpcd/hadoopexam6Solved")
```

Step 5: Check whether file has been created or not.

```
hdfs dfs -ls hdpcd/hadoopexam6Solved
hdfs dfs -cat hdpcd/hadoopexam6Solved/part-00000
```

Exercise-7: You have been given below code snippet.

```
val a = sc.parallelize(List("dog", "cat", "owl", "gnu", "ant"), 2)
val b = sc.parallelize(1 to a.count.toInt, 2)
val c = a.zip(b)
```

operation1

Write a correct code snippet for operation1 which will produce desired output, shown below.

```
Array[(String, Int)] = Array((owl,3), (gnu,4), (dog,1), (cat,2), (ant,5))
```

Solution:

```
c.sortByKey(false).collect
```

sortByKey [Ordered]: This function sorts the input RDD's data and stores it in a new RDD. The output RDD is a shuffled RDD because it stores data that is output by a reducer

which has been shuffled. The implementation of this function is actually very clever. First, it uses a range partitioner to partition the data in ranges within the shuffled RDD.

Then it sorts these ranges individually with mapPartitions using standard sort mechanisms.

Exercise-8: You have been given following product.csv file

product.csv (Create this file in hdfs)

```
productID,productCode,name,quantity,price
1001,PEN,Pen Red,5000,1.23
1002,PEN,Pen Blue,8000,1.25
1003,PEN,Pen Black,2000,1.25
1004,PEC,Pencil 2B,10000,0.48
1005,PEC,Pencil 2H,8000,0.49
1006,PEC,Pencil HB,0,9999.99
```

Now accomplish following activities.

- Using SparkSQL select all the records from file.
- Select all the products name and quantity having quantity < 2000
- Select name and price of the product having code as 'PEN'
- Select all the products , which name starts with PENCIL
- select all products which "name" begins with 'P', followed by any two characters, followed by space, followed by zero or more characters

Solution:

Step 1:

```
hdfs dfs -mkdir sparksql1
hdfs dfs -put product.csv sparksql1/
```

Step 2: Now in spark shell

// Import Spark SQL data types and Row.

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

// load the data into a new RDD

```
val products = sc.textFile("sparksql1/product.csv")
```

// Return the first element in this RDD

```
products.first()
```

//define the schema using a case class

```
case class Product(productid: Integer, code: String, name: String, quantity:Integer , price: Float)
```

// create an RDD of Product objects

```
val prdRDD = products.map(_.split(",")).map(p => Product(p(0).toInt,p(1),p(2),p(3).toInt,p(4).toFloat))
prdRDD.first()
prdRDD.count()
```

// change ebay RDD of Product objects to a DataFrame

```
val prdDF = prdRDD.toDF()
```

// register the DataFrame as a temp table

```
prdDF.registerTempTable("products")
```

// Select data from table

```
val results = spark.sql( """SELECT * FROM products """ )
```

// display dataframe in a tabular format

```
results.show()
val results = spark.sql( """SELECT name, quantity FROM products WHERE quantity <= 2000 """ )
```

```
results.show()
val results = spark.sql( """SELECT name, price FROM products WHERE code = 'PEN' """ )
results.show()
val results = spark.sql( """SELECT name, price FROM products WHERE upper(name) LIKE 'PENCIL%' """ )
results.show()
```

// "name" begins with 'P', followed by any two characters, // followed by space, followed by zero or more characters

```
val results = spark.sql( """SELECT name, price FROM products WHERE name LIKE 'P__ %' """ )
results.show()
```

Exercise-9: In Continuation of previous question, please accomplish following activities.

- Select all the records with quantity >= 5000 and name starts with 'Pen'
- Select all the records with quantity >= 5000, price is less than 1.24 and name starts with 'Pen'
- Select all the records witch does not have quantity >= 5000 and name does not start with 'Pen'
- Select all the products which name is 'Pen Red', 'Pen Black'
- Select all the products which has price BETWEEN 1.0 AND 2.0 AND quantity BETWEEN 1000 AND 2000.

Solution:

```
val results = spark.sql( """SELECT * FROM products WHERE quantity >= 5000 AND name LIKE 'Pen %' """ )

results.show()

val results = spark.sql( """SELECT * FROM products WHERE quantity >= 5000 AND price < 1.24 AND name LIKE
'Pen %' """ )

results.show()

val results = spark.sql( """SELECT * FROM products WHERE NOT (quantity >= 5000 AND name LIKE 'Pen %') """ )

results.show()
val results = spark.sql( """SELECT * FROM products WHERE name IN ('Pen Red', 'Pen Black') """ )

results.show()

val results = spark.sql( """SELECT * FROM products WHERE (price BETWEEN 1.0 AND 2.0) AND (quantity BETWEEN
1000 AND 2000) """ )

results.show()
```

Exercise-10: In Continuation of previous question, please accomplish following activities.

- Select all the products which has product code as null
- Select all the products, whose name starts with Pen and results should be order by Price descending order.
- Select all the products, whose name starts with Pen and results should be order by Price descending order and quantity ascending order.
- Select top 2 products by price
- Complete all the queries given in solution

Solution:

```
val results = spark.sql( """SELECT * FROM products WHERE code IS NULL """ )
results.show()


val results = spark.sql( """SELECT * FROM products WHERE code = NULL """ )
results.show()


val results = spark.sql( """SELECT * FROM products WHERE name LIKE 'Pen %' ORDER BY price DESC """ )
results.show()


val results = spark.sql( """SELECT * FROM products WHERE name LIKE 'Pen %' ORDER BY price DESC, quantity """ )
results.show()


val results = spark.sql( """SELECT * FROM products ORDER BY price LIMIT 2 """ )
results.show()


-- Define aliases to be used as display names
val results = spark.sql( """SELECT productID AS ID, code AS Code, name AS Description, price AS `Unit Price` FROM
products  ORDER BY ID""" )
results.show()


val results = spark.sql( """SELECT CONCAT(code, ' - ', name) AS `Product Description`, price FROM products""" )
results.show()
```

```
//Check "`" this is from the ~ key.
val results = spark.sql( """SELECT DISTINCT price AS `Distinct Price` FROM products""" )
results.show()


val results = spark.sql( """SELECT DISTINCT price, name FROM products""" )
results.show()


val results = spark.sql( """SELECT * FROM products ORDER BY code, productID""" )
results.show()


val results = spark.sql( """SELECT COUNT(*) AS `Count` FROM products""" )
results.show()


val results = spark.sql( """SELECT code, COUNT(*) FROM products GROUP BY code""" )
results.show()


val results = spark.sql( """SELECT code, COUNT(*) AS count FROM products  GROUP BY code  ORDER BY count
DESC""" )
results.show()


val results = spark.sql( """SELECT MAX(price), MIN(price), AVG(price), STD(price), SUM(quantity)       FROM
products""" )
results.show()


val results = spark.sql( """SELECT code, MAX(price) AS `Highest Price`, MIN(price) AS `Lowest Price`
     FROM products
     GROUP BY code""" )
results.show()


val results = spark.sql( """SELECT code, MAX(price), MIN(price),
          CAST(AVG(price) AS DECIMAL(7,2)) AS `Average`,
          CAST(STD(price) AS DECIMAL(7,2)) AS `Std Dev`,
          SUM(quantity)
     FROM products
     GROUP BY code""" )
results.show()


val results = spark.sql( """SELECT code AS `Product Code`,
       COUNT(*) AS `Count`,
       CAST(AVG(price) AS DECIMAL(7,2)) AS `Average`
     FROM products
     GROUP BY code
     HAVING Count >=3""" )
results.show()


val results = spark.sql( """SELECT
```

```
    code,
    MAX(price),
    MIN(price),
    CAST(AVG(price) AS DECIMAL(7,2)) AS `Average`,
    SUM(quantity)
  FROM products
  GROUP BY code
  WITH ROLLUP""" )

results.show()
```

Chpater-6: FAQ for HDPSCD Spark Certifications

Question-1: What is the Major change done in Hortonworks Spark (HDPSCD) in latest certifications?

Answer: There are following things have been changed recently

- **Platform**: Exam would be conducted on HDP 3 platform instead of HDP 2, and also Ambari 2.7 would be used.
- **Syllabus**: Two new sections are added or updated
    - Spark SQL
    - Structured Streaming

  Because Spark 2.x major focus is on the Spark SQL, which uses its own new version of Catalyst Optimizer to optimize the execution plan as well as its own Data types would be used. Hence, in the HDPSCD-Spark exam there is also focus shift towards the DataFrame and Structured Streaming.
- **Structured Streaming**: Previously in Spark for solving real time stream data problem we were using DStream solution, which was quite complex for developer as well as underline engine was different. And Structured Stream was built to make it developer friendly as well as underline execution engine is same as DataFrame/DataSet uses. Hence, the same code

(transformation, Aggregation, filtering etc) you can use on both Streaming and static batch data processing.

Question-2: Is Hortonworks test is on Hortonworks Enterprise platform or on the Apache Spark?

Answer: Hortonworks is conducting this exam on the HDP platform (this is a proprietary of the Hortonworks), however the concepts and assessment tested are based on the Apache Spark only. If you know the Spark then you just need to know how to access the data stored in HDFS, how to submit the job on YARN etc. Only few things additionally you should know, other than Apache Spark. So we don't consider too much difference because of platform. Rather, we say it make much easier to work with.

Question-3: Should I consider or preferred Scala based Spark certification only, because Spark is written using Scala and heard that Spark Scala is faster that PySpark?

Answer: You should select certification based on your programming language skill. If you are from Java/Scala background then go for Scala based Spark certification. And if you know Python programming language than go for Python Spark certification.

With regards to performance: that was the case on older version of Spark where Scala Spark was better performant than PySpark. But in Spark 2.x this is not the case (Because of their Optimizer), whatever programming language you use either Java/Scala/Python/R all are same on performance. Only exception to this is User Defined Function.

Question-4: What is the name of current version of Spark certification?

Answer: Hortonworks had changed and updated their Spark certification and now new name for the exam is HDPSCD (Hortonworks Spark Certified developer)

Question-5: What is the duration of the exam and number of questions?

Answer: Exam is for 2 hrs long and total number of Hands on questions or assessment you would be getting is around 7-10. Usually we have seen learners are getting total task 7.

Question-6: What all are the topics asked in the Hortonworks Spark certification exam?

Answer: There are mainly three major section based on which questions are asked.

- Core Spark
- Spark SQL
- Structured Streaming

We have already discussed each topic of the syllabus in previous chapter.

Question-7: I see RDD mentioned in the syllabus, so they are still part of the certification, even using Spark 2.x?

Answer: Syllabus mentioned for the Spark certification is very clear. And it is given in detail what they will be asking in the exam. And we expect quite a good amount of questions based on the RDD programming task. Reason RDD is still in focus, is because whether your use Spark1.x or Spark 2.x their underline processing engine works on RDD only. Hence, concepts of RDD must be cleared. And

you wanted to apply some custom optimization, wanted to do performance tuning etc. then you should know how RDD works. Even if you are using distributed shared variables like Broadcast variable and Accumulators then you will be using RDD.

You can convert your DataFrame to RDD, as well as RDD to DataFrame with simple API. Hence, you must have a good experience with the Spark RDD programming as well. And it is expected you know while working with the Spark.

Question-8: In the syllabus they have mentioned Streaming, so questions would be asked from Dstream or Structured Streaming?

Answer: Question would be based on the Structured Streaming. Hence, you should have a very good understanding about the different concepts of the Structured streaming.

Question-9: Is there any book is available for preparing Spark SQL, which can help in certification preparation as well?

Answer: HadoopExam has a very popular book for the Spark SQL which include 35 Hands On exercises as well. This is the link for accessing the same

Question-10: What about Spark Machine Learning Questions?

Answer: No, they are not part of this certification exam. This exam mainly focusing for the Spark Development and should be attempted the professionals who are working with Data processing like.

- Data Engineer
- Data Programmer
- Data Analyst
- Data Scientist
- Software Developer and Programmer

Question-11: I have some feedback and information about the Spark certification which needs to be updated here, for the benefits of the other learners?

Answer: Feedback is always welcome, this book and most of our material is being updated based on the feedback we receive from our learners. You can provide your feedback by sending an email on the hadoopexam@gmail.com or admin@hadoopexam.com

Question-12: Do we expect any question related to GraphFrame in the certification exam?

Answer: No, although GraphFrame is depend on the DataFrame and uses the same execution engine as used by the SparkSQL. But as of now in this certification we don't see any question is being asked on the GraphFrame or data processing using Graphs.

Question-13: Why Spark technology in so much news?

Answer: From the Apache it is one of the most actively worked framework. In recent years BigData, Real time Data processing, Artificial Intelligence and many other things pushed high. And all this need a processing engine which can process such things efficiently. Even Hadoop Mapreduce which become suddenly popular, is replaced by Spark computation engine. There are almost more than 1000 contributors on the open source platform.

After Spark 2.0, it is very easy to learn. Its API is very intuitive as well if you are good at SQL queries then it make it much easier for you to learn. If you are a programmer than DataFrame/Dataset API would help you a lot for working with the Spark.

There are many organizations who had pushed Spark applications in the production. Which proves the quality and reliability of the Spark framework.

Companies already having Hadoop cluster do not have to create separate Spark cluster. They can use their existing framework for the running Spark jobs on the same cluster. Whether it is written using Java/Scala/Python or R language.

Always having new technologies knowledge will give you the opportunity to draw more salary. And less chance of job loss. You can switch your career and Spark is one of them for sure.

Question-14: I have good knowledge of Spark, and almost 3+ years' experience working with Spark, why should I go for certification?

Answer: There is a myth in IT industry that certification does not help in career. This is not at all true. Having certification certainly helps in following ways

- You will know all the hidden features of a technology. If you go for certification
- It shows your career focus
- While resume shortlisting, it is given priority (Because first shortlisting is done by recruitment team, they don't have enough knowledge about technology. Hence, they look for your credentials in the resume).
- First impression on the interviewer.
- Interviewer will focus on things which you have written in resume.
- You will be categorized in separate category.
- It will give confidence during the interview and while working in the organization.
- So, avoid all the people who have -ve thinking about learning. Learning can never be costly and time wasting (universal truth).
- Certainly, and additional feather in your hat.
- There are many other latent benefits for doing certification.

Question-15: Do you give priority to specific vendor?

Answer: No, we don't give priority to any vendor. It varies based on many factors.

- Like if you wanted to get certified in both Hadoop and Spark then go for Cloudera Hadoop and Spark certification. And you have to have knowledge how to use Cloudera platform.
- If you are working on MapR platform then you can go for MapR Spark certification. Even other advantage is that MapR Spark certification is not as lengthy as Databricks Spark

certification. You can prepare for MapR Spark certification quite less time. You can see pros and cons that Databricks is more involved with the Spark and really tough one among the all Spark certification.

- Hortonworks Spark certification: This is again Hands on certification for the Spark. And have limited syllabus and specific objectives are given. Recently updated to include and support Spark 2.x version on the Hortonworks HDP platform.

Question-16: I don't know both Scala and Python then which programming language you would recommend?

Answer: It is very tricky question to answer. We recommend learn both the programming language. These are beautiful language to work upon. But based on the following career path you can choose respective programming language.

Scala:

- Java programmer should go for this
- If you want to become Data Engineer than go for this
- If you want to work on Data Cleaning and collecting Data than go for this.
- If you already know Java/Scala than go for this

Python:

- If you know Python than go for PySpark.
- If you are on Business Analytics profile go for PySpark
- I want to become Data Scientist, you can use either PySpark or Scala Spark

It should not be considered based on the fact that Spark is written in Scala, so I should give preference to Spark Scala. Not at all true after Spark 2.x version.

Question-17: What is the current passing score?

Answer: Current passing score is 75%, so if you get 7 tasks then you have to complete at least 5 to pass the exam. This exam is not at all tough, it is more about how much you have practiced. We have seen our learners are able to complete all 7 assessments in 90 mins, which is 30 min less than over all given time.

Question-18: What is the fee for the Hortonworks certification and how many attempts we can have?

Answer: Currently fee is $250, and you can have only 1 attempt. If you want to re-attempt then you have to pay fee again. (We are hoping, you don't have to re-attempt the exam)

Question-19: Any other or particular sections you want me to focus?

Answer: These all are the common area and you must keep in mind

- You will not get too many questions from RDD programming but for sure 2 to 4 questions you will be getting on RDD.

- You must know the partitioning and shuffling concepts, how to avoid shuffling. What is narrow transformation?
- How to read data formats like Parquet, ORC, CSV, JSON, XML and AVRO. You must know various options for reading the data using DataFrameReader object. What all options are available for reading such data?
- Once you read and process the data, how would save this data like in HDFS file system. There are various options and syntax, you must know them.
- For reading and writing assume you will get 6-8 questions in total.
- Major updates have been done for following two components in Spark 2.x
    - Project Tungsten
    - Catalyst Optimizer

Question-20: During the certification preparation, I am also preparing for the interview, can you please let me know, is there any material for the same?

Answer: yes, we do have interview preparation material for the Spark. Which you can get it from here. This is part of our premium and pro subscription.

Chpater-7: All Other Spark Certifications

- Databricks Spark Certifications
- Cloudera Hadoop & Spark Developer Certifications
- Hortonworks HDPSCD-Spark Certifications
- MapR Spark Developer Certifications

Databricks Certifications

Databricks is a company which was founded by Spark developer and this is the company which is actively involved with the Apache Spark development with the latest releases. Hence, if you want to go for Latest version of Spark Certifications than you can consider Spark certifications from Databricks. There are currently two certifications which are provided by Databricks one is in Scala and other is in Python programming language, however this exam is conducted on the Databricks own platform, so you may slightly different way of using the Spark. However, overall Spark API remain same. In both Scala and Python certification Syllabus remain same, but it is more of, for the learners who are comfortable in specific language. Current version of the Databricks Spark certification is tested with the Spark 2.x version, below is the syllabus outline for the Databricks Spark certification.

Spark Architecture Components: Candidates are expected to be familiar with the following architectural components and their relationship to each other:

- Driver
- Executor
- Cores/Slots/Threads
- Partitions

Spark Execution : Candidates are expected to be familiar with Spark's execution model and the breakdown between the different elements:

- Jobs
- Stages
- Tasks

Spark Concepts: Candidates are expected to be familiar with the following concepts:

- Caching
- Shuffling
- Partitioning
- Wide vs. Narrow Transformations
- DataFrame Transformations vs. Actions vs. Operations
- High-level Cluster Configuration

DataFrames API: Candidates are expected to have a command of the following APIs.

SparkContext: Candidates are expected to know how to use the SparkContext to control basic configuration settings such as spark.sql.shuffle.partitions.

SparkSession : Candidates are expected to know how to:

- Create a DataFrame/Dataset from a collection (e.g. list or set)
- Create a DataFrame for a range of numbers
- Access the DataFrameReaders
- Register User Defined Functions (UDFs).

DataFrameReader: Candidates are expected to know how to:

- Read data for the "core" data formats (CSV, JSON, JDBC, ORC, Parquet, text and tables)

- How to configure options for specific formats
- How to read data from non-core formats using format() and load()
- How to specify a DDL-formatted schema
- How to construct and specify a schema using the StructType classes

DataFrameWriter: Candidates are expected to know how to:

- Write data to the "core" data formats (csv, json, jdbc, orc, parquet, text and tables)
- Overwriting existing files
- How to configure options for specific formats
- How to write a data source to 1 single file or N separate files
- How to write partitioned data
- How to bucket data by a given set of columns

DataFrame (Dataset):

- Have a working understanding of every action such as take(), collect(), and foreach()
- Have a working understanding of the various transformations and how they work such as producing a distinct set, filtering data, repartitioning and coalescing, performing joins and unions as well as producing aggregates
- Know how to cache data, specifically to disk, memory or both
- Know how to uncache previously cached data
- Converting a DataFrame to a global or temp view.
- Applying hints

Row & Column: Candidates are expected to know how to work with row and columns to successfully extract data from a DataFrame

Spark SQL Functions : When instructed what to do, candidates are expected to be able to employ the multitude of Spark SQL functions. Examples include, but are not limited to:

- Aggregate functions: getting the first or last item from an array or computing the min and max values of a column.
- Collection functions: testing if an array contains a value, exploding or flattening data.
- Date time functions: parsing strings into timestamps or formatting timestamps into strings
- Math functions: computing the cosign, floor or log of a number
- Misc functions: converting a value to crc32, md5, sha1 or sha2
- Non-aggregate functions: creating an array, testing if a column is null, not-null, nan, etc
- Sorting functions: sorting data in descending order, ascending order, and sorting with proper null handling
- String functions: applying a provided regular expression, trimming string and extracting substrings.
- UDF functions: employing a UDF function.
- Window functions: computing the rank or dense rank.

How to prepare for Databricks Spark Certifications?

To prepare for any technology certification of your interest you should consider 2-3 months' timeline. If you get the well organized and focused material. If you don't get the organized and

certification focus material it would be difficult to prepare the exam and timeline can increate upto 9-12 moths and by the time many things would change with respect to certifications. Hence, to prepare you for the Databricks Spark Developer certifications in Scala and Python you should consider below preparation material.

This material is available on http://www.HadoopExam.com

- **Trainings**: If you are not familiar and having average experience of the Spark frameworks than we recommend below trainings which will help you prepare for these certifications
    - Apache Spark Professional (Include 2.x latest Version setup) Training with Hands On Lab
    - Spark 2.X SQL (Using Scala) Professional Training with Hands On Sessions
    - Scala Professional Training with HandsOn Session
    - Python Professional Training with Hands-on Sessions
    - PySpark Structured Streaming professional training

- **Practice Questions and Answers**: To save time and focused approach for Spark certifications you should go through the below certification material for Databricks Spark certifications.

- CRT020 : Databricks Certified Associate Developer Apache Spark 2.4 with Scala 2.11 : Assessment Certification (Newly launched & Active)
- CRT020 : Databricks Certified Associate Developer for Apache Spark 2.4 with Python 3.0 - Assessment Certification (Newly launched & Active)

Cloudera Hadoop and Spark Developer Certifications:

Cloudera is a pioneer for Hadoop framework and they have lot of frameworks for BigData paradigm. Cloudera provide one of the mostly used Hadoop Framework and known as CDH (Cloudera Hadoop Distribution). CDH is bundle of various big data software and one of them is Spark. Cloudera also focuses on Spark for data processing rather than traditional MapReduce frameworks. Hence, they are also delivering Spark software as part of their CDH distribution. Cloudera has various certifications for Hadoop and BigData professionals. For the Spark developer one of the most popular certifications since last 2 yrs. is been this CCA175 (Cloudera Hadoop and Spark Developer certification)

In this certification 30%-40% focus is on Spark and remaining part is Hadoop Data Processing.

How to prepare for CCA175?

On http://www.HadoopExam.com this is the certification preparation material which is most subscribed among many top 10 certifications. HadoopExam provide a combined package for preparing CCA175 which include below three products.

- Spark Professional Training. with Hanson Session
- Hadoop Professional Training with Hands-on Session
- CCA175 Spark and Hadoop Developer Certifications (Includes 111 Solved Scenarios and Complimentary videos for selected solutions)

MapR Spark Certifications: The *MapR Certified Spark v2.1 Developer* credential proves that you have ability to use Spark to work with large datasets to perform analytics on streaming data. This credential measures your understanding of the Spark API to perform basic machine learning or SQL tasks on a given datasets.

This material is available on http://www.HadoopExam.com

- **Trainings**: If you are not familiar and having average experience of the Spark frameworks than we recommend below trainings which will help you prepare for these certifications
    - Apache Spark Professional (Include 2.x latest Version setup) Training with Hands On Lab
    - Spark 2.X SQL (Using Scala) Professional Training with Hands On Sessions
    - Scala Professional Training with HandsOn Session
    - Scala Professional Training with HandsOn Session

Practice Questions and Answers: To save time and focused approach for Spark certifications you should go through the below certification material for Databricks Spark certifications

- About MapR MCSD: MapR® Certified Spark Developer: Total 220+ Solved Questions: Recently updated based on learners feedback.

Where and How to get Databricks Spark CRT020 Certification Sample Questions

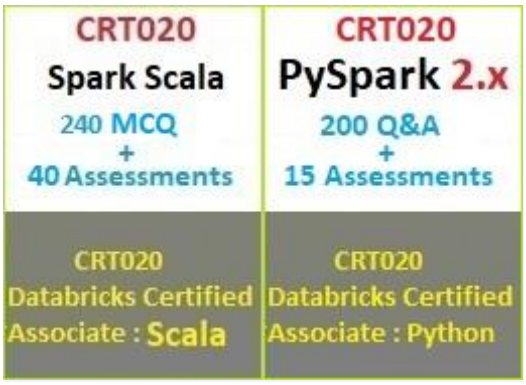
There are various Spark certification exams available and this particular one "CRT020 : Databricks Certified Associate Developer for Apache Spark 2.4 and Scala 2.11 - Assessment Certification Exam " is the latest available Spark exam from the Databricks. This certification became popular in very short span of time and within the launch on HadoopExam.com , more

than 100 learners have subscribed in a week. This prove that, how popular is this certification exam. And also this is based on the Databricks Enterprise version of the platform.

Even it uses the Databricks Enterprise version but its underline engine is same as Apache Spark, hence, the same code you can run on the Apache Spark as well as Databricks Spark platform.

However, it is recommended that you practice very well before you appear in the real exam. Because without practice, you would not be able to complete the exam on time. CRT020 exam is divided in two major section as below.

- Multiple Choice Questions (Get access to all 240 Multiple Choice Questions from Here Scala , PySpark)
- Assessment (Hands On Section) : Get access to all 40+ assessment Questions and Answer (Including Videos) Scala or PySpark



If you want to check the Sample Questions and Answer then use the below link or watch the below video to understand more.

- Scala : http://hadoopexam.com/spark/databricks/SparkScalaCRT020DatabricksAssessment.html
- PySpark : http://hadoopexam.com/spark/databricks/PySparkCRT020DatabricksAssessment.html
- Sample Assessment PySpark: http://learn.hadoopexam.com/PySparkCRT020/SampleAssessment/index.html
- Sample Assement Scala : http://learn.hadoopexam.com/SparkScalaCRT020/SampleAssessment/index.html

- Multiple Choice
  : http://learn.hadoopexam.com/SparkScalaCRT020/Sample/index.html

How you should prepare for CRT020 Spark Scala/Python (Databricks) Certification Exam?

Databricks is the leader for Apache Spark technology, they support the open source version of Apache Spark framework.

Based on the open source Apache Spark, Databricks created enterprise version of Spark Framework. And this newly created framework also work on the Cloud platform like AWS, Azure, Google cloud etc.

Since last few years Databricks platform became very popular because they are capable of deploying Spark in the production env. Enterprise or companies who all are using Databricks platform in production or planning to have in production. They are looking for Databricks certified professionals. Databricks has following two popular certifications as of today. They might come more in future for different solutions like Machine Learning, Graph and Structure Streaming etc. Let's go through below two links for the currently available certifications.

1. CRT020 : Databricks Certified Associate Developer Apache Spark 2.4 with Scala 2.11 : Assessment Certification (Newly launched & Active)
2. CRT020 : Databricks Certified Associate Developer for Apache Spark 2.4 with Python 3.0 - Assessment Certification (Newly launched & Active)

Both the above certification exam has the same pattern. Only difference is the, which programming language you prefer to give the exam.

Exam format : In each certification exam there are two sections as below

1. Multiple choice questions and answers (which include single and multiple correct answers, fill in the blanks questions and answers etc.)
2. Assessment Exam : You need to write complete solution for given problem statements. Also link would be provided for downloading or accessing the data.

However, it is not clearly mentioned that how many questions they would be asking in each sections. However, HadoopExam.com experience shows that there would be around 40 multiple choice questions and 5-10 assessment exercises would be given where difficulty level would increases Question by Question, Same is provided in HadoopExam online Spark Certification Simulator. It is clearly mentioned that the exam would be 3 hrs long and include both the above section. Hence, please note that

- In multiple choice 40-50 questions and answers would be covered. In that they would be asking various concepts, internal architecture, API and SQL functions based questions.
- Around 5-10 assessment questions would be asked, in this you would be given problem statement for each questions and you need to write or implement the solutions either using PySpark or Spark Scala. Based on the version of exam you selected.
- You need to write problem solution in online version of Databricks Enterprise platform.
- How the Scoring would be done? : Databricks have not mentioned, whether you need to pass separately each exam section or aggregate score from both the section would be considered. What HadoopExam.com experience again says here that you need to score 75% marks in each section at least so that your overall aggregated score remain 75% as well and you can clear the exam. Whether Databricks consider individual section or aggregated marks.

Timeline for CRT020 Spark Certification preparation: Preparations and timeline depend on the how good you are in Spark technology as well as your strength in Scala and Python programming language. As per HadoopExam.com experience following timeline you can consider for preparing this certifications, if you spend 2-3 hrs 5 days a week (we are giving you two days off in a week, it could be any two days).

- **6 month** : If you are completely new to Spark.
- **3-4 month** : If you know one of the programming language like Java, Scala, or Python etc.
- **1-2 month** : If you already know Spark technology.

Above timeline is not perfect these are derived based on HadoopExam.com previous experience with other certifications.

How to prepare for CRT020 Spark Certification: to prepare for the Spark certification you need to have right material, and also you need to properly planned and properly drafted material, which can save your lot of time. Otherwise, you would be going for material here and there and lose lot of time and it may take much longer to complete the exam even without having full confidence in the real exam.

Also, remember if things are not properly planned and drafted or organized, It does not matter how good you are in Spark.

To make your life simple and easy for the Spark CRT020 certification preparation HadoopExam.com have created cool material. You should consider the following material for preparing Spark Certification

1. CRT020 : Databricks Certified Associate Developer Apache Spark 2.4 with Scala 2.11 : Assessment Certification (Newly launched & Active) : Include 200+ multiple choice questions and more than 30 assessments.
2. CRT020 : Databricks Certified Associate Developer for Apache Spark 2.4 with Python 3.0 - Assessment Certification (Newly launched & Active) : Include 200+ multiple choice questions and more than 30 assessments.
3. Apache Spark Professional Training with Hands On Lab Sessions (Active)
4. Spark 2.X SQL (Using Scala) Professional Training with Hands On Sessions
5. PySpark 2.X (Using Python) Professional Training with Hands On Sessions
6. Scala Professional Training with HandsOn Session
7. Python Professional Training with HandsOn Session

All the required questions come with the full explanation of the questions and answer. To justify the correctness of the questions and answers.

- It covers the entire syllabus for both Python and Scala version of certification exam. You can attempt their questions and answers as many times as you want.
- All multiple-choice question and answer, you can access from any device where browsers are supported like Desktop, Macbook/IOS, iPhone, mobile, tablet etc.
- There is no separate installations are required.
- Most of our learners are happy that because while travelling or during free time they can access the certification preparation material as well as interview questions audio cum video book.
- You can check some sample questions and answers as below
    - **Check Sample Assessment Paper (Scala)**
    - **Check Multiple Choice Sample Paper(Scala)**
    - **Check Sample Assessment Paper (Python)**
    - **Check Multiple Choice Sample Paper (Python)**

- **Video Cum Audio Book : Spark 2.x Interview Preparations (Total 185+ Interview Questions) : Video + Audio + PDF**

More detail on assessment exam: HadoopExam.com give capability to you for accessing problem statement and assessment solutions which can be accessed from mobile and tablet and that you can understand the same in detail. Once you understand the problem statement, then in the next tab, you would be given instructions to access or download the data which you need to use for solving the problem statement.

Videos: Possibly for selected assessment would have videos as well as, in the HadoopExam.com would explain the entire problem statements and its solution. However, it is not guaranteed that each assessment would be having the videos.

Assessment Solution: We are providing step by step solution for the given problem in multiple steps. Each step would be written with the detailed comments as well. So that you can easily understand what is being done in the solution. Check the below video for more detail.

Training: HadoopExam.com has very popular training for Apache Spark, Spark SQL, Structured Streaming in Python and Scala. As well as interview Questions Audio cum video books. These all are On-Demand training access which you can access anytime anywhere using mobile, desktop, MacBook, iPhone etc. Check all below.

Interview Preparation: By going through certification exam and training, your ultimate target is to join the companies which are using these new platforms or if you are already working in the organization then you are looking for vertical growth or increase on pay package and salary. Hence, HadoopExam.com prepared almost 185+ Interview Questions and answers which you can access in these two formats EBook and Video cum Audio Book format. These material if you want to read you

can read, you want to watch you can watch and if you want to listen then you can listen as well anytime-anywhere. If internet access is available to you. Check more detail as below

Why Cloudera CCA175 Hadoop and Spark developer certification is more popular?

No doubt that Cloudera is one of the Pioneer and leader for the big data technology. And Cloudera really created the market for big data and also did very good job for Hadoop framework.

Similarly in case of Hadoop and Spark certification CCA175 Cloudera not only evaluate the Spark technology but also evaluate the Hadoop skill. And you have to solve all the given problem on Cloudera cloud-based platform.

The reason why most of the companies are looking for professional with Cloudera CCA175 Hadoop and Spark developer certification, because they have already deployed Cloudera enterprise platform in the production environment, companies in the domain like investment banks healthcare IT companies, retail E-Commerce companies, airline and travel platform, start-up which are working on data science research projects as well as machine learning solutions.

There's another reason, like Hadoop can be easily deployed on cloud platform for example AWS, Azure etc.

There is another feather recently added by merging Cloudera and Hortonworks together to lead the big data technology world.

The reason why Cloudera always remain leader because it continuously accepts new technology and update their platform very frequently compared to any other provider . For example Cloudera have adopted recent version of spark as well. They have very good support for hive, pig, OoZie, and their own develop solution Impala which can run much faster than hive.

These are the only few reason and there are much more which made Cloudera platform very popular in the big data world.

So in CCA175 exam Cloudera evaluate your skills based on 8 to 10 problems solutions which you need to solve using Hadoop Hive pig and spark (you can use either 1.x or 2.x version of Spark, its upto you). Cloudera is really not worried what technology you use to solve a problem but rather

they want problem should be solved efficiently. Whether you use map-reduce, Hive, Impala or shell script for cleaning up the data. There would be at least three to four exercises on Apache spark, in that they would give you already implemented some solution in the form of template and you need to fill in the remaining part using the functional programming either in Python or Scala. It is clearly said by the Cloudera that questions template would not be given in both python and Scala language for the assessment. It is up to you whether you want to write entire program your own or you want to use existing skeleton (template) provided by the Cloudera during the exam. Hence it is expected you are very good on the Apache Spark Core as well as Spark SQL at least.

This exam has higher value because it evaluates both the Hadoop and Spark in single certification exam. Complete name of the exam is CCA175 Spark and Hadoop developer. Where CCA means Cloudera Certified Associate. You can check the entire syllabus here on this page where we have provided the detailed description as well. If you have been given 10 problem statement in the real exam then at least seven problem statement you have to solve completely to clear the exam. We have seen most of our learners have scored around 9 to 10 problem solutions comfortably in the given time slot. We got the feedback that without practicing all the material provided by the HadoopExam.com you would not be able to complete the exam on time. As well as learners are able to complete the exam with the correct solutions and we are happy to share with you the same things. HadoopExam.com is providing Cloudera certification preparation material since last 6 years and our technical team had good expertise on that.

Currently the cost for this certification exam is 295 dollar, but we have seen sometime Cloudera give good discount on the fee as well or some companies have purchased coupon in bulk.

Other than this, what we have seen Impala and Hive mostly used to solve problem. For Spark they provide the skeleton for the problem scenario and you can use either Scala or python to solve the given problem. Skeleton would be provided only in one of the language like Python or Scala. If you know Scala and Cloudera provided skeleton in Scala, you may use this skeleton to complete the program, it may help you save the time during the exam. However, it is not mandatory that you use the skeleton provided rather you can completely write entire program from scratch.

As most of learners use the HadoopExam.com preparation material and with this practice material they are comfortably completing the exam on time or before and scoring around 9 to 10 questions perfectly.

Now how do you get this preparation material for CCA175 certification? Use the below link to get respective material
Use this 90+ solved scenario for Cloudera CCA 175 Spark and Hadoop developer certification.

1. In this material you would be provided instruction to setup the environment for practicing all scenarios.
2. Instruction would be provided to get the data for practicing the questions.
3. Step by step solution is provided for each problem statement.

4. for selected and complicated problem scenarios, videos are also provided and trainer would explain problem and solution in detail.
5. If you want to understand more on that then watch the below video.

Cloudera CCA175, Hortonworks HDPCD & Databricks CRT020 Certification Exam

There are various Spark Certification available as below and these very popular IT certification

1. Databricks Spark Certification for Developer CRT020 in Scala or Python.
2. Cloudera CCA175 Hadoop & Spark Developer in either Scala or Python.
3. Hortonworks Spark (HDPCD) certification in scala & Python
4. MapR Spark Developer certification in Scala only.

All above certification has equal value, respective certification importance increases when based on the company in which you are working or giving interview and which platform this company is using.

For example, if company has the Cloudera platform already deployed in production then CCA175 certification exam would be more useful and certainly have more value addition then other company certifications. Similarly, if company had deployed Databricks platform in production then your Databricks Spark CRT020 certification would have more values.

How should I compare these Company Certification with training institutes certifications?

Training institutes certification does not have that high importance because these institutes does not take any protected exam like above company and as soon as you pay the high amount of fee, you are entitled to get the certificate of training attended. It does not matter whether student learned in the training or attended training or not. Institute really does not evaluate the candidate's expertise. Hence, company does not consider them until and unless you have valid certification from global companies like Cloudera, Databricks, Hortonworks, MapR etc. Even experience says, students are more grilled during the interview if they write local training institutes training.

About Global certification from above companies

- Cloudera CCA175 => Spark (Either Scala or Python) + Hadoop
- Databricks Spark CRT020 => Spark Core + Spark SQL in Scala or Python
- HDPCD Spark => Spark Core + Spark SQL + Spark Structured Streaming (Either Python or Scala)

As you can see in above Cloudera CCA175 certification both Hadoop and Spark would be accessed. Hence, their syllabus would be covering two wider domains. However, the level of exam difficulty is moderate and not very tough. Most of our students have score either 9 or 10 questions correctly in the real exam. They have prepared using HadoopExam CCA175 certification simulator.

In case of Databricks main focus is on Core Spark and Spark SQL. But in-depth knowledge would be evaluated. Like you should be aware minor details about the Spark framework, internal architecture of Spark framework, what all stages your job goes through when submitted on the multi node cluster etc. You must be knowing how the distributed data sharing works and much more. And all the concepts are evaluated using the multiple-choice questions and answer in the real exam. Databricks CRT020 has two sections one is multiple choice exam and other is assessment (check the entire syllabus at this page).

Databricks have not explicitly mentioned that how many questions they would be asking in the assessment, but HadoopExam experience says that in total 4-8 problem statements would be given and you need to solve them on the Databricks platform (Check here, how to solve such problem). And through this assessment exam Databricks would be accessing your knowledge, experience and efficiency in the Spark Core API, Spark SQL, SQL functions, how to use user defined functions, how to cache, un-cache and checkpoint the DataFrame and Datasets etc. So mainly your programming expertise would be checked. Total exam would be for 3 Hrs.

However, Databricks had not mentioned or given any specific detail for number of questions, passing score, how much time is allotted for each section or overall time is 3 hours for both the section. (As on when we have update, we would share with you, so please be in touch with http://hadoopexam.com

Regarding passing score our past experience suggests that , you score 75% marks at least in each section. So that your average score would also be 75% and you can easily clear the certification exam.

Similarly, HDPCD Spark is also fully HandsOn certification exam and you need to solve all the given problem on Hortonworks Cloud based platform. It is purely Spark framework certification exam but they are not limited to the Core Spark rather they want you to know core Spark API, Spark SQL, Spark Structured Streaming. However, you can write solution based on your choice of the programming language like Python or Scala.