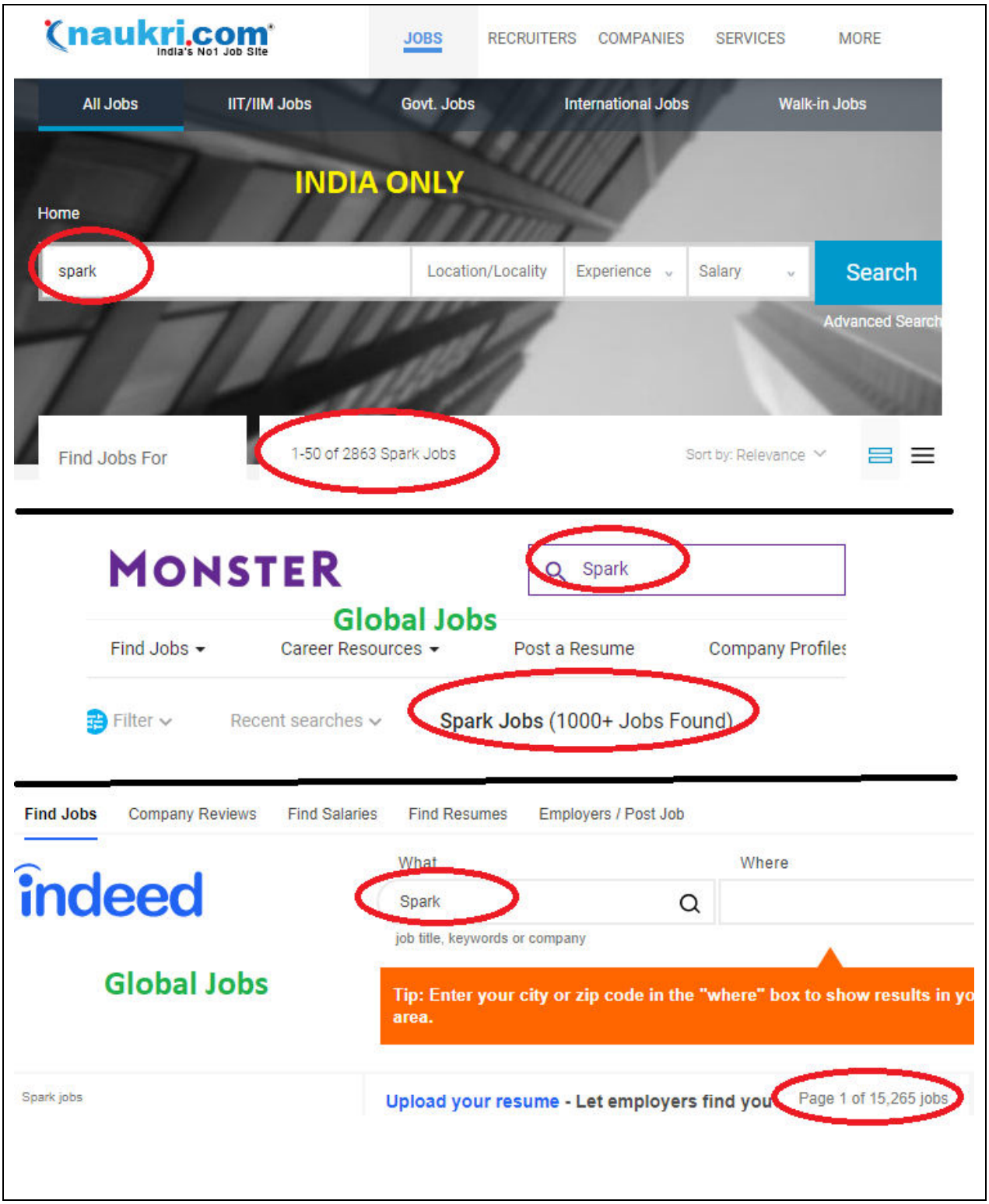
MAPR SPARK CERTIFICATION PREPARATION GUIDE

By HadoopExam.com

About Spark and Its Demand:

One of the top trending technology since last 3-4 years. One of the highest subscriber on HadoopExam for Spark trainings and certifications, with lot of enquires for existing as well as new upcoming products pertinent to Spark technology. These all proves Spark capabilities and its demand. From below screen as of 14-April-2018, you can see number of available jobs in India as well as globally. Note that, this does not include Hadoop, if you have knowledge of Hadoop and Cloud computing like AWS, Azure, Google Cloud then it's even great.

- India Alone: 2863 open positions
- Globally : 15000+ open positions



Spark become popular because of its speed of data processing as well as very much compatibility with the already popular Hadoop framework. It can process data from RDBMS, Local File System, Amazon S3, HDFS, NoSQL e.g. Cassandra, HBase. Reason it can process data with high speed because, it use less disk

(unless required or you configured to use disk) and most of the computation is done in memory (Disk I/O is a killer for computation performance). Remember Spark is not a storage engine, it's a computation engine. Spark framework is written in Scala programming language, but its API can support Python, Java, Scala, and R languages. Beauty with the Spark is not only its computation engine but various abstraction layers created over the Spark core engine. Let's look few of the popular abstractions, which make Spark more user/programmer friendly. This framework is not only used by Data engineer but recently demand has increased among data analysts, statisticians, mathematicians and data scientists. If you are really looking for becoming data scientist (Sexiest job of 21st century) and Spark framework must be in your resume and being certified (**Hindi : Sone pe suhaga**) is really give you an edge. Our technical team work in industry and they know, what value it carries, if you are Spark certified.

- Core Spark: This is the core API, on which other Spark frameworks are created. RDD and PairRDD are the core of the Spark engine. Yes, you need to be very well versed with the RDD API to use and understand how Spark framework works under hood.
- SparkSQL: This is the layer created and recently optimized for better performance. It supports SQL queries like Joins, Unions, Having clause, where clause all are supported and SQL engine will convert these SQL queries in Spark Jobs so that they can be efficiently executed.
- Spark Streaming: Continuous data processing. It is very challenging to process real time data. Spark gives a framework to process continuous stream of data like Twitter feed, Stock market real time trading data etc.
- GraphX: Data problems which has relationship can be very well solved using GraphX framework.
- Machine Learning: Many of the renowned Machine Learning algorithm already implemented by the Spark Framework. You just need to know, which API is there for your required algorithm.

Who should learn Spark?

As I mentioned previously Spark is becoming popular framework and almost every industry started using it from retail, finance, aviation, insurance, you name it, they are certainly started using it. So who should start learning it, it's a framework which has API in 4 languages as well as many abstractions are created for end user and programmer. Most of the below professionals are using it.

- Java Programmer/Developer
- Python Programmer/Developer
- Data Analysts
- Data Scientists
- Data Engineer
- People who are working with R language

About Spark Certifications:

Let's discuss about various certifications available for Apache Spark. Below are the popular Spark certifications of 2018. They are in existence since last 3 to 4 years and many thousands of learners already get certified using http://www.HadoopExam.com certification preparation material.

1. Cloudera Spark and Hadoop Developer (CCA175) : Hands On Certification, Download 111 Solved Scenario from here
2. Hortonworks Spark Developer (HDPCD : Spark) : Hands On Certification, download 65 solved scenario from here.
3. MapR Certified Spark Developer (MCSD) : Multiple choice Questions and Answer, Download 220+ questions from here
4. Oreilly Spark Developer (Retired, and no more available) : Multiple choice certification
5. Databricks Spark Developer : Multiple choice certification

Our main focus in this guide is for MapR Certified Spark Developer, which recently started becoming popular. There are preferences among the learners we found. Some learners prefer HandsOn certification exam (They assume they are more valuable, but that is not true). Some people consider to appear in multiple choice based questions (They feel this is easier than HandsOn exam, again this is not true).

HandsOn Exam: It has limited syllabus to cover and format of the questions are well known, so after doing good practice with these HandsOn questions provided by http://www.HadoopExam.com you can easily clear your exam (most of our learners are scoring 9/10 or 10/10, if you are good at time management). Yes, for HandsOn exam you really have to do practical's and long questions and answer.

Multiple Choice Questions: Certification exam based on this would certainly have wider syllabus and they can ask as minute detail as possible. So thorough understanding of the Spark concepts is required. It is not enough that you are able to write program in Spark, so you will be able to clear this certification exam. This exam has focus on testing both your understanding about the Spark framework, how can you optimize computations, how efficiently you can configure to run on multi-node cluster. Reading and writing data from and to multiple types of storage and data formats. We will discuss about MCSD (MapR Spark Certification Developer) in detail in next section of this guide. So what you should know, I am letting you know in brief

- Must be well versed with the API for programming questions (more than 70 % questions will be on this)
- You should know, how Spark framework can be configured, so that it can be efficiently utilized.
- What parameters, you should set for your individual Spark job is performance friendly.
- What happens, with the Jobs once you submit the job on cluster?
- How would you debug your submitted jobs?
- How would you monitor the Spark job submitted by you as well as entire Spark cluster?
- How to write efficient SQL queries, which can avoid network I/O by reducing data shuffling.
- What all are file formats are available which are supported by Spark Framework.
- How to process continuous stream of data.
- How Spark streaming works and how small batches can be created in Spark streaming.
- Basic knowledge of Machine Learning
- How to use Machine Learning (MLib) library etc.

So you can see multiple choice exam is equally tough to clear. So it require good understanding of Spark framework as well as its underline workings. You must consider these 220 questions and Spark

Professional training provided by HadoopExam to clear MCSD and very helpful while working in real-time environment.

MCSD Certification Syllabus in detail: As per the MapR study guide they have 7 topics which they will be covering in real certification exam. We will discuss about each topic and what you need to focus.

1. Load and Inspect Data: This activity is always needs to be done whenever you are going to work with huge volume of data. Before starting any processing on the data you need to load this data and it may be possible your data is not in the format, in which you want it. So you will wrangle with this data and create as per your further processing need. Hence, you should know how to load data from HDFS, S3, RDBMS, NOSQL DB and Local File System. Also you should also aware how to convert it into the format, which you need. Let's discuss each subtopic under this.
    a. Creating RDD: You should know, what exactly the RDD is. How can you create RDD from various data format e.g. JSON, CSV, Sequence File, Parquet File and Avro files etc. How do you create RDD from Java collection?
    b. Transformations **:** This is one type of operation you apply on the RDD and which create another RDD. Remember RDD's are immutable, so you always have to create new RDD from existing RDD, if you want to format/transform your data. There are various transformation API is available and the most commonly used are map(), flatMap(), reduceByKey()
    c. Actions on RDD**:** Transformation helps you convert your RDD from one format to another format or filter the data. But to get the result or do some calculations on the RDD you need to apply action. Here, very important concepts come into the picture is Transformations are lazy and only evaluated until you call action on the RDD.
    d. Caching and Persisting the RDD**:** Understand the difference between RDD caches and persist. What all options are available for that and how and which API you will be using for caching and persisting RDD. When and in which scenario it is useful and optimal to cache the RDD.
    e. Actions v/s Transformation**:** As we have already discussed Transformations are lazy and only evaluated once actions are called on the RDD. You will not get direct questions, what is the difference between Transformation and actions, but they will tweak some coding questions and you should be able to answer that question by understanding the concepts of transformation and actions.
2. Spark Application: There are mainly two ways by which you can execute your Spark program, one is through interactive shell and other is by creating Spark Applications (Bundled Jars in case of Java and Scala). Interactive shell is good for implementing prototypes and checking basic functionality of the Spark Applications, finally in production you need to run Spark Applications. So you need to know, how to create Spark applications and what all things are needed. Let's discuss each subtopic in detail.
    a. MapReduce jobs on YARN: Yes, in this case you need to know basics of Hadoop framework and MapReduce algorithm. If you are aware Hadoop framework was

created based on two main concepts MapReduce (Compute) and HDFS (Storage). Both are distributed. So whatever, you can implement using Spark Framework API, most of that can be written in MapReduce as well. However, people are not writing MapReduce too much now, because it requires lot of complex coding. However, concepts wise you should still aware how MapReduce works on Hadoop Framework. Now another challenge here is MapReduce part of Hadoop had evolved and new Framework has been created which is known as MapReduce 2.0 or YARN. YARN (Yet another resource negotiator) is another framework, which not only supports MapReduce algorithm but others as well like Spark Jobs for parallel processing. So in the exam you may be asked questions like how the Spark jobs works when it is submitted on the YARN framework. I don't expect too many questions on this.

b. SparkContext**:** Whenever you need to submit an application to Spark, you need to know all the detail about entire cluster and its environment. So you need access to SparkContext. Spark provide the ability to create SparkContext instance for your application and you can use that instance during your application/job execution to get the details about Spark cluster. In interactive shell Spark provides you pre-created SparkContext object and that can be referred as an object "sc"
c. Main Application: As I have mentioned previously, that you should be able to create Spark Applications using main method. If you are able to write Spark code in interactive shell, then it is not challenging. You should know basic concepts of Scala how a class can be created and how to define main () method in it. This is more about structuring the long source code in small units and then build a Jar using either Maven or Scala build tool (SBT). However, I am not expecting too many questions on that. You should have basic concepts clear how to create Spark Applications, using classes, objects and main method. The main point here is how you create SparkContext for your application. Because SparkContext is main entry point for the Spark Application in Spark framework. As you know in the Spark interactive shell object of SparkContext is already available as an object which is referred by a value/variable "sc" (Do you know the difference between value and variable in Scala?)
d. Difference between Interactive shell and application: Nope, you may not be asked what the difference between these two. Rather as I mentioned what is available in Spark shell there are other variables also available in Spark shell which are HiveContext, SQLContext, and SparkContext etc. But when you submit your application, you should be able to create your own these objects. How to create these all objects. Very important point to remember is that when you will be creating HiveContext object, it is not necessary to have Hive and Hadoop framework in place. You can create HiveContext objects with Hadoop framework, even you should prefer to use HiveContext rather than SQLContext for writing SQL query.
e. Application Run Mode**:** There are various cluster manager on which Spark can run its application. There are four main cluster manager Local, Standalone, YARN and Mesos. I would say among all these 4, YARN is most popular one. What exactly are the differences when you use one of these cluster manager and when you should consider in which scenario? I would say Local and standalone mode are more for testing and prototyping. You should be aware how YARN framework works and what happens

when Spark job is submitted on YARN cluster. How the Spark Executor, Tasks, worker etc. co-ordinated.

3. PairRDD: It's simple to understand, if you have ever worked with the MapReduce algorithm (you see many of the things are existing but new frameworks are created for optimization, performance and usability perspective). PairRDD is an RDD of key-value pair, it is a tuple in Scala with only two values like (value1, value2). So here value1 is a key and value2 would become value. And PairRDD represents such data. Beautiful API is available to work with the PairRDD (paired data). You don't have to write so much code, like you to with the MapReduce programming, just few lines of code with the PairRDD API it could accomplish many things in distributed manner. Most of the places you can see word count application as an example to show you, that how 50 lines of MapReduce code can be written using Spark PairRDD with just 5 to 6 lines to accomplish the same thing. Let's discuss each topic in detail about PairRDD for MCSD certification perspective.
    a. Creating PairRDD**:** First of all you need to understand the concepts of the PairRDD and how it can be used. Once you understand this then learn how to convert simple RDD into PairRDD, when you load the data how it can be directly converted into PairRDD. Similar to simple RDD, PairRDD also have actions and transformations. How serialization affect while working with keys and values of the PairRDD. Yes, you will get many questions around simple RDD and PairRDD, both combined will cover at-least 30-35% questions. Most of the questions will be asked with code snippet and sample data.
    b. PairRDD transformations**:** This is something difficult to grasp, as reduce() function in case of simple RDD is an action but in case of PairRDD, reduceByKey() is a transformation. So, how do you differentiate between transformation and actions? Concept remain same transformation will return new RDD and action return the expected results. When actions are applied in RDD, then transformation will be lazily evaluated.
    c. reduceByKey and groupByKey **:** Certainly, you will get questions based on these two methods of PairRDD. So it is expected you understand how this functions work with the data in RDD, where is and when data shuffling, what is the issue with the Network I/O etc. Which should be consider in which situation.
    d. PairRDD functions**:** There are various functions available in PairRDD for transformations and actions. You need to understand, how to use them for example join, leftjoin, rightjoin, union etc. What is the role of the key, when join operation is applied. Basically, you should have good exposure to PairRDD API. You will get 4-5 questions with both actions and transformations of PairRDD.
    e. Action on PairRDD**:** Similarly you should have good experience with the use of PairRDD actions. Certainly 2-3 questions from here.
    f. Partitioning**:** Very important concepts, you should be able to create partitioning of data. How the partitioning affect your data processing, how overall performance is impacted. Like having more partitions will give you highest parallelism, but when you need to shuffle data, then partitioning is kill for performance. Yes, 2-3 questions on this section as well. Different types of available partitioner.

4. DataFrame: One of the beautiful abstractions created for the Spark programmer, you can convert an RDD to DataFrame and DataFrame can be converted into RDD. Once you have DataFrame created (You have to use Scala case classes to assign schema to your RDD). You can use various SQL queries on this DataFrame. It make your overall work very simple. You should be able to convert an RDD to DataFrame and vice-versa. How to register this DataFrames as a temporary tables/views, so that you can query them. Other advantage of using DataFrame is, optimizes your queries for execution perspective.
    a. Creating Data Frame**:** As mentioned you need to convert existing RDD into DataFrame, using reflections. So it means, you should understand what the case classes in Scala are and how to use them for converting RDD to DataFrame. Again, you need to save back query output in a local or HDFS file system.
    b. Running Queries**:** Once you created DataFrame, write a code which register DataFrame as a temp table. Once, registered as a temp table, apply various queries. If you know basic SQL like select, where, joins, unions, subtract etc. then it would be quite easy for you to work with the Spark DataFrame. You not only have SQL's to be applied on DataFrame, but it comes with the convenient API as well, so many people prefer to use API, rather than SQL. So you should be well versed with both the use cases.
    c. UDF: If you are familiar with the SQL, then you can see there are some functions you can use like count(), max(), min() etc. These are commonly used functions, and whenever you have some custom requirement which are not simply solved by existing functions, you can create your own custom functions. Once, you create this functions, you have to register them, so that you can use in your SQL queries. This functions are known as User Defined Functions, before using them you have to register them.
    d. Re-partitioning: We have already discussed about the partitioning in Spark and its impact on performance. Sometime, looking the impact you want re-partition DataFrame, how can you do that. What is the impact of re-partitioning? You must be aware about this.
5. Monitoring: Once, you defined your application/ or Spark Job and submitted to Spark cluster for execution. That is not good enough. You need to know, what is happening with the Job you submitted to Spark cluster. Is cluster has enough resources to run my job, the configuration parameters I have been using for the job I submitted is good enough? This all required real-time monitoring. There are various ways, by which you can monitor your application. Let's discuss all the ways by which we can monitor our submitted application for performance perspective.
    a. Stage, Tasks and Jobs: Do you know the concept of job, stage and tasks. When you submit the jobs. It is very important for any Spark Job you submit, you are able to calculate how many stages and tasks would be there. Reason, you can gauge the performance of you submitted application to Spark cluster. This way you can conceptually understand, how your application will perform on Spark cluster.
    b. Spark Web UI: Spark framework comes with its own web UI, where you can monitor each individual Spark job. Once you submit the job to Spark, it will create a link specific to your job on this web UI and by clicking on it, you can monitor your individual application as well, you can see, if you have defined any counters in that application.

c. Performance Tuning: You may be thinking you have written Spark application very well for solving your problem, but somehow, you not only look the correct API, but other factors are also very important like. Is API you used is correct for parallelism, or having less data shuffling or it will spill data over disk. Have you done caching of RDD on right place? How the current cluster configuration is affecting your application. Once you have understood correctly, you should be able to tweak individual configuration parameter for your application. Learn concepts of data locality etc. Data locality will reduce the data shuffling.

6. Spark Streaming: This is one of the reason, Apache Spark framework become suddenly popular in BigData world. There were many existing framework for processing continuous stream of data but they were not convenient and not giving the expected performance as well and huge volume of data is another challenge. Same event/message and data should not be processed again. Another beautiful feature in Spark streaming is Window operation on continuous data. This entire complexity is very well implemented in Spark framework. So, I can say, you need to know your business logic and DStream API to work with. All the complexity of handing this stream on fault-tolerant manner is the Spark framework responsibility. Let's discuss the topics, which will be asked in the exam.
    a. Spark Streaming Architecture: You have to understand the concept of micro batching in case of Spark streaming. Which will create DStream objects, this DStream can also have transformation and action API. What is the minimum window size you can create in Spark streaming, how to apply Spark functions on this DStream, how to do analytics on continuous stream of data, it's very challenging.
    b. DStream API: Live data is sent as small batches, Dstream is nothing but batches of RDDs. Dstream represents a continuous stream of data, namely a sequence of RDDs. RDD is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. (Learn Spark Streaming in detail from here) A Discretized Stream (DStream), the basic abstraction in Spark Streaming, is a continuous sequence of RDDs (of the same type) representing a continuous stream of data. DStreams can either be created from live data (such as, data from TCP sockets, Kafka, Flume, etc.) using a StreamingContext or it can be generated by transforming existing DStreams using operations such as map, window and reduceByKeyAndWindow. While a Spark Streaming program is running, each DStream periodically generates a RDD, either from live data or by transforming the RDD generated by a parent DStream. This class contains the basic operations available on all DStreams, such as map, filter and window. In addition, PairDStreamFunctions contains operations available only on DStreams of key-value pairs, such as groupByKeyAndWindow and join. These operations are automatically available on any DStream of pairs (e.g., DStream [(Int, Int)] through implicit conversions. DStreams internally is characterized by a few basic properties: - A list of other DStreams that the DStream depends on - A time interval at which the DStream generates an RDD - A function that is used to generate an RDD after each time interval
    c. Dstream stateful operations: In Apache Spark 1.6, it had been improved support for stateful stream processing with a new API, mapWithState. The new API has built-in support for the common patterns that previously required hand-coding and

optimization when using updateStateByKey (e.g. session's timeouts). As a result, mapWithState can provide up to 10x higher performance when compared to updateStateByKey. One of the most powerful features of Spark Streaming is the simple API for stateful stream processing. Programmers only have to specify the structure of the state and the logic to update it, and Spark Streaming takes care of distributing the state in the cluster, managing it, transparently recovering from failures, and giving end-to-end fault-tolerance guarantees. While the existing DStream operation updateStateByKey allows users to perform such stateful computations, with the new mapWithState operation we have made it easier for users to express their logic and get up to 10x higher performance.

d. Actions on Dstream**:** Similarly you can apply actions on the RDD, you can apply on the Dstream as well. Once, you apply the action it will initiate the computation on the Dstream data. Similarly you can save the Dstream data using saveAsTextFile and some other available API. Example of Action is foreach method of Dstream.
e. Window operations: We have already discussed little about the windowing function in Apache Spark. Window function means like you're receiving continuous stream of data every 100 millisecond. Now you want to apply some calculations on all the data you have received in each 1 second (Like calculating average bid price on HadoopExam stock ticker). What you will do in this case, you define window size as either in time format like on each second, which will consider last 10 micro-batches. Or you can create window based on number of batches as well. You need to understand in depth. And there are various API methods, which you can use countByWindow, reduceByWindow, countByValueAndWindow etc.
f. Fault-tolerance and process only once: When you are processing continuous stream of data, then it is very challenging you should not lose any single bit of data, as well as no data should be processed by twice. Yes, you are not worried about this because Spark Streaming framework handle this. But you need to know, how Spark take care of this problem. You will get question based on this criteria as well.

7. Advanced Spark and MLib: As you are involved with the Spark framework, you will start using some advanced concepts. These advanced things are very helpful for writing efficient Spark program. Like you know, what is the broadcast and accumulators in Spark and how to use them? Can RDD be directly broadcasted, so that lookup data will be available on each node of the Spark cluster? No, you need to convert RDD into Scala collection before broadcast (you will only get to know, if you have done some hands-on exercises, else it is not that easy). Spark developer had create Machine Learning library and well known and mostly used algorithms of Machine Learning are already implemented. You should have some basic knowledge of Machine Learning and able to find which algorithm fall under supervised learning and which in un-supervised learning. What exactly the difference between these two, what is clustering, classification and understanding of recommendation engine etc. Let's discuss each topic in little more detail. It is not expected that you are a Machine Learning expert.
    a. Broadcast variable: When you have two sets of data, one is very small (can fit into memory) and another is very huge and want to join both the data. Hence, to get the performance, what you will be doing broadcasting small set of data. So it would already be available on each node. You must know, which data can be broadcasted and which

should not. Here, only factor is size of the data. Smaller size data should always be broadcasted.

b. Accumulators: When you wanted to do some counting (like counters in MapReduce framework), you can use accumulators provided by Spark framework. Accumulators can only be updated on the worker node of the Spark cluster and final aggregated values can be received at the driver. The most important concepts in case of Accumulator, is that you should be able to use it in actions only. Nobody, is going to stop you using accumulators in transformation, but it will not give the correct results in case of node crash, or same function is executed twice on different node for performance etc.
c. Supervised v/s unsupervised learning**:** You must know the difference between supervised and unsupervised learning like supervised learning has some set of test data which can be used to supervise the output. Something is there to supervise the outcome. But in case of unsupervised learning there is no such test data and results.
d. Classification: You will be given some data and needs to classify them in pre-defined classes e.g. email Spam filter.
e. Clustering: grouping of data, and you don't know initially what all groups will be created once this algorithms are executed on the data.
f. Recommendation: If you have purchased or visited some products on website, based on that you will be recommended similar products.
g. MLib: You need to identify, which library and API, you will be using for particular Machine Learning Algorithm.

Sample Questions with Answer and Explanations: Let's see some of the questions we have created for the MapR Certified Spark Developer (This questions are taken from the http://www.HadoopExam.com MCSD certification simulator, where currently we are providing in total 220 questions with answers. Please visit this link for more detail about MCSD certification.

Question 1: You have been given following code written in Scala and Spark. Below is the content for IBM.csv file

IBM,101,20150112
Google,400,20150112
IBM,107,20150113
Apple,230,20150112

Now you have written following code, in interactive shell

val myRDD = sc.textFile("data.csv")
val splittedRDD = myRDD.map(_.split(","))
val value = splittedRDD.map(x=>x[0]).XXXXX.count()

Please replace XXXXX ith correct function, which will produce output value as 3

1. map(x=>len(x))

2. distinct()
3. filter(X=>X.contains("IBM")
4. No function is needed , it will be a redundant call

Correct Answer : 2 Exp : Steps are follow
1. Load data.csv file in myRDD (each line as a record in RDD)
2. Split the RDD based on comma (,). So it will create list of lists [[IBM,101,20150112].....]
3. Now select only first element of each row
4. Using the distict() function, we can select only distinct stock name
5. And once we have distict stock name, we just call a count() function on it. So it can generate desired output.

Question 2: You have been given following code written in Scala and Spark. Below is the content for IBM.csv file

IBM,101,20150112
Google,400,20150112
IBM,107,20150113
Apple,230,20150112

Now you have written following code, in interactive shell
val myRDD = sc.textFile("data.csv")
val splittedRDD = myRDD.map(_.split(","))
val value = splittedRDD.map(x=>(x[0],1)).XXXXX.collect()

Please replace XXXXX ith correct function, which will produce output value as
Array((IBM,2),(Google,1),(Apple,1))

1. reduceByKey((X,Y)=> X+Y)
2. reduce((X,Y)=> X+Y)
3. groupBy((X,Y)=> X+Y)
4. countBy((X,Y)=> X+Y)

Correct Answer: 1 Exp: reduceByKey(func), is a function, which applies func function on each tuple for each key. So each key will produce the sum of count. It is similar to word count example.

Question 3: You have been given following code written in Scala and Spark Below is the content for IBM.csv file

IBM,101,20150112
Google,400,20150112
IBM,107,20150113
Apple,230,20150112

Now you have written following code, in interactive shell

val myRDD = sc.textFile("data.csv")
val splittedRDD = myRDD.map(_.split(","))
val distinctRDD = splittedRDD.map(x=>(x[0],1)).distinct()
val priceDataRDD = myRDD.map(x=>(x[1]))

In above program, which of the following RDD should be cached.

1. myRDD
2. splittedRDD
3. distinctRDD
4. priceDataRDD

Correct Answer: 1 Exp: If we are using same RDD, again and again then it is advisable to cache or persist the same. Cached RDD has already been computed and the data is already in memory. We can reuse this RDD without using any additional compute or memory resources.

Question 4: What is an RDD?

1. When you start a cluster, it will create pool of RDDs to store in Memory data during application execution.
2. It is a pool of predefined storage like 64MB or 128MB but created lazily, whenever applications are submitted.
3. They are collections of objects which are distributed across nodes in a cluster, but created by applications code itself.
4. It is a container for the source code for your application and will be distributed on all the nodes of the cluster.

Correct Answer: 3 Exp: RDD is primary abstraction in Apache Spark, An RDD represents collection on objects that is distributed across nodes in a
cluster. Most of the computation or operation you will be doing will be performed on RDD.

Question 5: Which of the following is a correct way to create an RDD, assume sc is an instance of SparkContext (API syntax are given not correct)

A. sc.textFile("Path to a text file")
B. sc.sequenceFile()
C. sc.hadoopRDD()
D. sc.hadoopFile()
E. sc.loadSequenceFile()

1. A,B,C
2. B,C,D
3. C,D,E
4. A,D,E
5. A,C,E

Correct Answer: 1 Exp: Apache Spark API supports various file format to be loaded as an RDD. For example
textFile() : It reads the file and will return an RDD that contains one record per line
wholeTextFile() : it helps you to read a directory containing multiple small text files and returns an RDD as (Key, FileContent). Filename as a key and
its content as a value
sequenceFile() : For reading sequence file
hadoopRDD() : to read hadoop format files

Question 6 : In unsupervised learning which statements correctly applies

1. It does not have a target variable
2. Instead of telling the machine Predict Y for our data X, we're asking What can you tell me about X
3. telling the machine Predict Y for our data X
4. 1 and 3
5. 1 and 2

Correct Answer: 5 Exp: In unsupervised learning we don't have a target variable as we did in classification and regression.
Instead of telling the machine Predict Y for our data X, we're asking What can you tell me about X? Things we ask the machine to tell us about
X may be What are the six best groups we can make out of X? or What three features occur together most frequently in X?

Tips & Tricks for Certification: Based on our experience talking with the learners who had already appeared in real exam, we have created some tips and tricks, which you need to know before appearing in real exam.

- Very less theoretical questions (Around 20% questions based on concept)
- 80% questions will be based on Code snippet and Sample data.
- No questions are being asked on GraphX as of now.
- **No direct question**: you need to know the underline concept to correctly answer the question.
- Quite complex questions.
- Question on using code snippet with map and flatMap functions
- Difference between supervised & un-supervised learning. Which one is unsupervised learning algorithm with below options?
    - Supervised Learning
    - Understand basics of Regression
        - Linear regression
        - logistic regression
    - Understand classification algorithms
    - Naive baise classifiers
    - SVM (Simple vector machine)
    - Random decision forest.
    - Unsupervised Learning
    - Dimension reduction.
    - PCA
    - SVD
    - K-means clustering
    - Difference between classification and clustering
- Maximum questions are from RDD : Around 17 Questions
- SparkSQL and DataFrame around : 14 Questions
- Spark Streaming 7 Questions
- Machine Learning 7 Questions

- PairRDD, Monitoring, Stage, Lineage : 10 Questions
- Broadcast variables and Accumulators : 3-5 Questions
- Partitioning and Re-partitioning : 7 Questions
- Understand ReduceByKey, GroupByKey and Reduce functions (Questions are certain from this)
- Configurations parameters related questions to improve the performance, what is the memory requirement for executor etc.
- They might give a practical scenario with sample data, with some cluster information like 10 nodes, 30 executors, and HDFS directory containing 100 files, which will be loaded by Spark Job. What all optimizations are possible, what is the memory needs to be configured, what is wrong with current configuration etc.
- RDD caching and persist questions (2-4 questions)
- Initial value of the Accumulator will be given and once job complete what will be the final value of the Accumulator.
- In a Spark Job, how many stages will be executed and how many will be skipped, based on RDD cache.
- Find possible number of partitions.
- Output format of the Spark Job.
- You will be given Code snippet and you need to select correct output from given question.
- Understand API method of Spark Streaming ReduceByKeyAndWindow.
- Practice and understand PairRDD functions like : groupByKey, reduceByKey, combineByKey
- Understand RDD API function like fold and reduce.
- Read little bit about MLib datatypes.
- Understand LabeledPoints
- Streaming window operations are very important.
- How fault tolerant is achieved is spark streaming.
- How back pressure is achieved in streaming?
- How to tune spark job
- Spark UI questions will be asked, but they are quite simple. Hence, just visit at least once Spark Web UI.
- Read DataFrame API methods.

Spark Version: Currently MapR Certification is using Spark 1.6 version. So be careful, you prepare for that specific version. (Let us know, if they have changed the exam with newer version, so our tech team look into that and inform you about product updates)

Spark Training: HadoopExam has created a Spark Training which covers core Spark to Advanced Spark almost all or more than 90% topics from this certifications. Hence, this is very useful training, you have to hardly spend 20-30 hours with this Hands-On training to become Apache Spark expert. Everything is explained in detail and very minute detail is covered. Please visit below link for Spark Scala Professional Training. This is one of the highest subscribed training on HadoopExam platform.

- Spark Professional training

Other Products:

HadoopExam has various products for BigData, Cloud Computing, Analytics, SAS, Programming and Certifications related. You can get entire product lists on home page only.

- All HadoopExam Products

Experience in Spark:

It is good, if you have experience in Spark before appearing in this certification MapR recommends that you should have 6-12 months experience at-least. If you don't have experience and not get an opportunity to work with Spark and want to enter in this high demanding technology that this training is very useful, which has hands on sessions included.

Successfully clearing the exam:

We hope once, you complete following products for preparing MapR Spark certifications, you will be able to clear real exam.

Other Spark Certifications:

Appendix: All Products available currently on HadoopExam platform

All Premium Training Access Annual Subscription (You will get early access to under development training) : Used By More than 20000 subscribers

Apache Spark Training & Certifications: Apache Spark is new and fastest data processing engine for Big Data world, after Hadoop it's becoming more popular in Industry (recently demand increased a lot). Now using power of Hadoop and Spark. Hence, data processing speed has dramatically increased. So if you wish to work in/with Big Data then Learning Spark is a must even for becoming data scientist., HadoopExam Learning Resources launched low cost material for in depth learning of Spark in the form of Spark Professional Training with Hands on practice sessions and helping you to get certified with most popular Apache Spark Certification conducted by Oreilly and Databricks only. So without delaying start preparing or prove your skills of Apache Spark, subscribe to our trainings and certification material with special discount of unbeatable price. You can request free updates as well, whenever it is done.

1. Apache Spark Professional Training with Hands On Lab Sessions
2. Oreilly Databricks Apache Spark Developer Certification Simulator
3. Hortonworks Spark Developer Certification
4. Cloudera CCA175 Hadoop and Spark Developer Certification

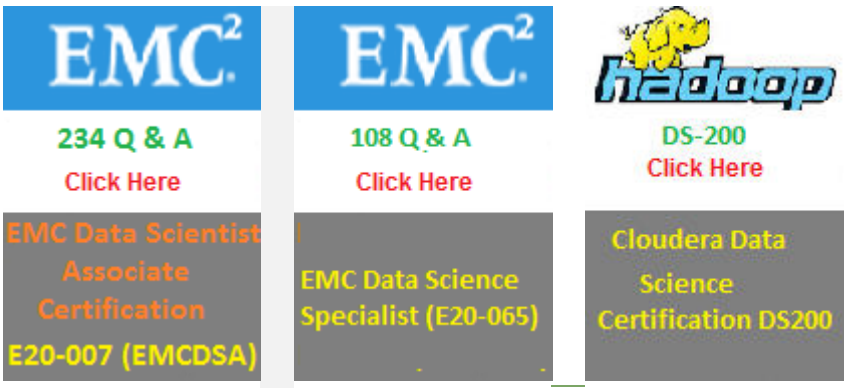
Cloudera® Certifications Preparation Kits and Trainings: Cloudera is a pioneer for Hadoop Big Data framework and they have grown a lot since last a decade. Cloudera® solutions is being used a lot in industry. They had also converted all their certification exam from multiple choice to Hands-on exam. HadoopExam was the first one, who launched Cloudera certification material 5 years back and since than we have also grown and keeping in pace with Cloudera new certifications. We also provide industry class training used by more than 10000 learners across the globe. Check all the products below for more detail.

1. CCA 175 : Cloudera® Hadoop & Spark Developer : 95 Solved Scenarios
2. CCA159: Cloudera® Data Analyst Certification : 73 Solved Scenarios
3. CCA131 : Cloudera Hadoop Administrator Certification : 92 Solved Scenarios
4. CCP:DE 575 : Cloudera Hadoop Data Engineer : 79 Solved Scenarios
5. CCA 500 : Hadoop Admin Certification : 250+ Practice Questions
6. Training : CDH : Cloudera Hadoop Admin Beginner Course-1 : 30 Training Modules
7. Hadoop Professional Training
8. HBase Professional Training
9. Cloudera Hadoop Developer (CCD410) Certification : Retired
10. Cloudera HBase Specialist (CCB400) Certification: Retired
11. Hadoop Package Deal

About Hortonworks® Training & Certifications: Hortonworks is one of the leader in providing Big Data solution through their own HDP platform. To check candidate's proficiency or skills for HDP platform they have various certification exams. HDPs most of the exam are Hands-on exam other than HCA (Hortonworks Certified Associate). All the exam aspirant has to solve given tasks on HDP cluster. In each exam there are approx. 10-12 problem scenario would be given and needs to be solved in 2 Hrs. Being an Hands-on exam, these certifications has high value in industry, because it require real hands on experience to solve given scenario. Hence to help you, HadoopExam is providing from scratch how to setup environment to practice scenarios. HadoopExam also provides the complementary videos, where we guide you how to solve problems and setup the environment. Currently we have following certification preparation material available.

1. HDPCD : Hadoop (HDP) No Java Certification : 74 Solved Scenarios
2. HDPCD-Spark : HDP Certified Developer : 65 Solved Scenarios
3. HDPCA : HDP Certified Administrator : 57 Solved Scenarios
4. Hortonworks Certification Package Deal

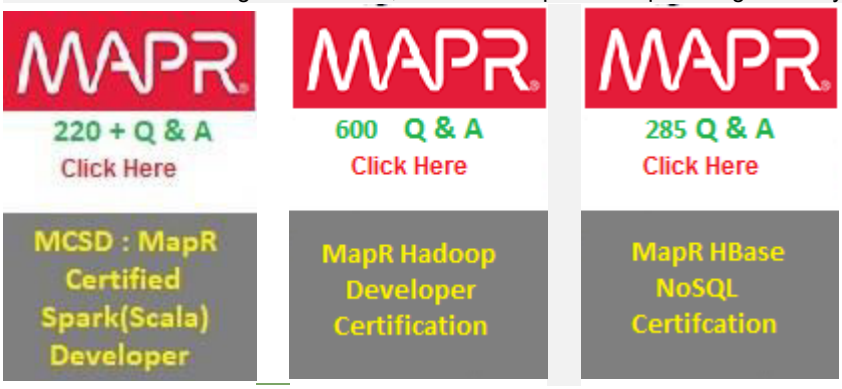
Data Science & Machine Learning: Data Science is one of the most demanding field, currently and we are providing following products to become a data scientist from one of the popular organization in the data world EMC



1. Data Science Certification EMC® E20-007 (Data Science Associate)
2. EMC® Data Science Specialist (E20-065)
3. Cloudera Data Scinece DS-200 (235 Questions + 150 Page Study Notes) : Retired

---

MapR® Training & Certifications : MapR is another most popular BigData solution provider based on Hadoop. These are the following certifications, which HadoopExam is providing currently.



1. MCSD : MapR Spark (Scala) Certified Developer
2. MapR Hadoop Developer Certification
3. MapR HBase NoSQL Certifcation
4. MapR Package Deal

AWS Training & Certifications : In the Cloud computing world , Amazon is a pioneer and most used Cloud Computing solutions. Currently there are following products are provided bt HadoopExam for the AWS trainings and certifications preparation. We have been providing this matrial since last approx 5 years and many 1000s of learners already using our material to grow in their career.

1. AWS Solution Architect Associate : Training
2. AWS Solution Architect Associate Certification Preparation
3. AWS Solution Architect Professional Certification Preparation
4. AWS Sysops Certification Preparation
5. AWS Developer Certification Preparation

IBM® BigData Architect : This is a multiple choice exam conducted by IBM for a BigData Architect. IBM also has Hadoop framework known as BigInsight and they will be asking Question based on BigInsight, however it is very similar to Hadoop only, because they are using Apache Hadoop framework only. As you know, IBM is the oldest and one of the matured software vendor and they have more penetration in the Industry, compare to any other BigData vendor. Hence, certifying yourself as a BigData Architect for IBM, ceratinly have high value in industry.

- IBM C2090-102: IBM Big Data Architect : Total 240 Questions : Highest number of Questions : 95% Questions with explanations

---

DataStax® Apache Cassandra Certification: This is a multiple choice exam conducted by DataStax for Apache Cassandra. DataStax is one of the leader in providing Apache Cassandra based solutions. Apache Cassandra is one of the most demanding and used NoSQL database across the industry. Cassandra has been used in Finance, HealthCare, Aviation, Retail, e-commerce and many more. It has proved itself with high degree of performance. However, it's a different database and RDBMS principals does not fit with Cassandra. You certainly need to learn Cassandra Data Modeling to design database perfectly and this certification is designed towards this only. And HadoopExam had put lot of effort to create this material to help in clearing this certification exam.

- Professional Certification Apache Cassandra(Datastax) : Total 207 Questions : Highest number of Questions : 95% Questions with explanations

---

SAS®: One of the most used commercial solutions for analytics, Data science, mathematical and statistical modeling. In analytics world no other solution is close to SAS. Its leader in its field and mostly used across industry. Below are the all products provided by HadoopExam.

1. SAS Base Certification Professional Training
2. SAS Base Programming Certification(A00-211)
3. SAS Certified Advanced Programmer for SAS 9 Credential
4. SAS Certified Statistical Business Analyst Using SAS 9: Regression and Modeling Credential
5. SAS Certified Platform Administrator 9 (A00-250) Certification Practice Questions
6. SAS Package Deal

HBase Training & Certifications: HBase is a NoSQL solution based on Hadoop framework. Hence, is very well compitible with the Hadoop based solution. You should certainly learn HBase, if you are wroking in BigData world using HadoopExam. Following are the products provided by HadoopExam for HBase.

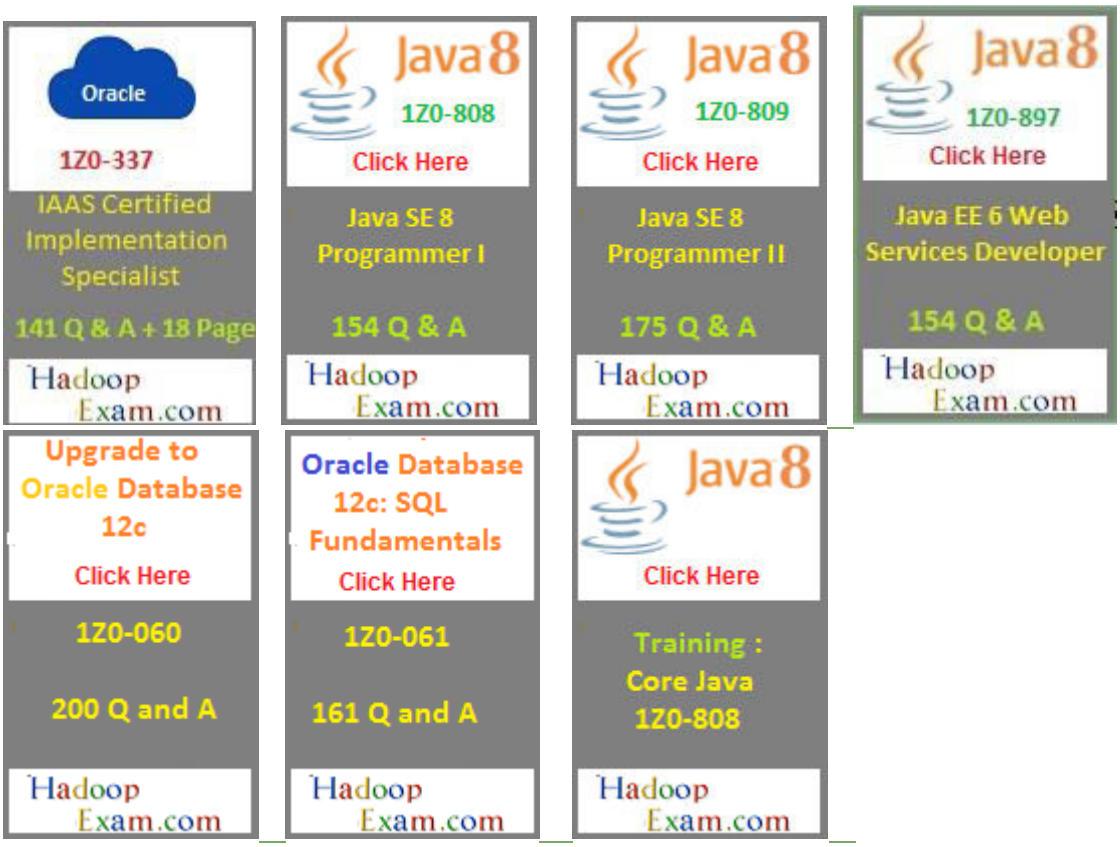
1. HBase professional Training with HandsOn Sessions
2. MapR HBase certification preparations

Microsoft® Azure: Microsoft Azure is another provider for Cloud computing solutions and also heavily used in the industry. If you are planning to make your career in Cloud computing than you should have very good understanding of the Microsoft Azure. Please find all the products and solution provided by HadoopExam for the Azure.

1. Microsoft Azure 70-532 Developing Azure Solution Certification
2. Microsoft Azure 70-533 Implementing Microsoft Azure Infrastructure Solutions

---

Oracle Cloud , Java and Other Programing Trainings and Certifications: There is no development without a programming skills. We provide trainings and certification material which will make you developer who can work in well developed IT industry, with the most demanding programming skills. So start learning Java, Scala, Python and complete its certifications as well. Please check all the available products below.



1. Oracle 1Z0-337 Oracle Oracle Infrastructure as a Service Certified Implementation Specialist
2. Full length HandsOn Step By Step Training for Java 1z0-808)
3. Scala Professional Trainings with HandsOn Session
4. Python Professional Trainings with HandsOn Session
5. Java SE-8 Programmer-1 (1z0-808) Certification
6. Java SE-8 Programmer-2 (1z0-809)
7. JAVA EE Web Services Developer (1z0-897)
8. Oracle® 1Z0-060 : Upgrade to Oracle Database 12c Administrator
9. Questions for Oracle 1Z0-061 : Oracle Database 12c: SQL Fundamentals