Hadoop and Spark Developer - CCA

CCA 175 Prep Plan

The key to any certification preparation is to have a proper plan, as the old saying says 'failing to plan is planning to fail'. In this blog I am going to show you one possible way you can prepare and obtain CCA 175 certification. My goal is to accomplish two things on this blog

1. Identify the technologies to learn in order to accomplish the certification goals
2. Create a realistic schedule that encompasses learning and practicing before appearing for the certification

Note: video walkthrough of this plan is available at [Click here]

Click here for the video version of this series. This takes you to the youtube playlist of videos.

Below table maps the required skill to technologies one needs to learn in order to solve problems during the certification exam. Remember, CCA 175 is a hands on exam, it is an open book exam but the only content you can access during the exam is api and official framework documentation. Hence, it is very important to gain a good level of comfort in using a set of hadoop eco system technologies, generic or specific frameworks and programming/query languages.

Skill Category	Skill Description	Technology To Use
Data Ingest	Import data from a MySQL database into HDFS using Sqoop	**Sqoop**
	Export data to a MySQL database from HDFS using Sqoop	**Sqoop**
	Change the delimiter and file format of data during import using Sqoop	**Sqoop**
	Ingest real-time and near-real-time streaming data into HDFS	**Flume or Spark Streaming**
	Process streaming data as it is loaded onto the cluster	**Flume or Spark Streaming**
	Load data into and out of HDFS using the Hadoop File System commands	**HDFS Command Line**
Transform,Stage and Store	Load RDD data from HDFS for use in Spark applications	**Spark RDD and Spark DF**
	Write the results from an RDD back into HDFS using Spark	**Spark RDD and Spark DF**
	Read and write files in a variety of file formats	**Spark RDD and Spark DF**
	Perform standard extract, transform, load (ETL) processes on data	**Spark RDD, Spark DF and Hive**
	Use metastore tables as an input source or an output sink for Spark applications	**Spark RDD, Spark DF, Spark SQL, Hive, Impala**
Data Analysis	Understand the fundamentals of querying datasets in Spark	**Spark RDD, Spark DF, Spark SQL**
	Filter data using Spark	**Spark RDD, Spark DF, Spark SQL**
	Write queries that calculate aggregate statistics	**Spark DF, Spark SQL, Hive and Impala**
	Join disparate datasets using Spark	**Spark RDD, Spark DF and Spark SQL**
	Produce ranked or sorted data	**Spark RDD, Spark DF, Spark SQL, Hive and Impala**
Configuration	Supply command-line options to change your application configuration, such as increasing available memory	**Spark Submit and options that can be used along with Spark Submit**

This essentially boils down to learning below tools, frameworks, libraries and technologies. Here are the pre-requisites before you start your learning journey and also for practicing these technologies.

- Basic knowledge of any programming language. If you have scala or python background then it makes it much more easier
- Good Understanding of what data and database means. Some knowledge of SQL querying also helps.
- Finally, the most import aspect of this practical learning is to have an environment. it may take hours or days for you to build a haoop environment with all these combination of technologies. Cloudera makes it easier for you by providing a quickstart VM that you can install on your machine. Please read the instructions carefully and watch some youtube videos on how to setup the quickstart VM for your practice. You can download the quick start VM here [Click HERE]

#	Technology	Languages	Description
1	**HDFS**	Unix like commands	The Hadoop Distributed File System (HDFS) offers a way to store large files across multiple machines.
2	**Sqoop**	Unix like commands with some SQL	Framework for bulk data transfer between HDFS and structured datastores as RDBMS.
3	**Spark**	Scala OR Python	It is a Data analytics cluster computing framework. Spark fits into the Hadoop open-source community, building on top of the Hadoop Distributed File System (HDFS). To its credit, Spark provides an easier to use alternative to Hadoop MapReduce and offers performance up to 10 times faster than previous generation systems like Hadoop MapReduce for certain applications.Spark is a

4	Spark RDD	Scala OR Python	RDD stands for Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.
5	Spark DF	Scala OR Python	A DataFrame is a distributed collection of data, which is organized into named columns. A DataFrame can be constructed from an array of different sources such as Hive tables, Structured Data files, external databases, or existing RDDs.
6	Spark Streaming	Scala OR Python	Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data
7	Spark SQL	Scala, Python and SQL	Spark SQL is a Spark module for structured data processing. Unlike the basic Spark RDD API, the interfaces provided by Spark SQL provide Spark with more information about the structure of both the data and the computation being performed. Internally, Spark SQL uses this extra information to perform extra optimizations.
8	Spark Submit	Unix like commands	it is a mechanism to run spark programs as applications by suppliying configurable parameters to optimize spark code execution
9	Flume	Unix like commands	Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.
10	Hive	SQL	Hive provides a SQL-like interface to data stored in HDP.
11	Spark Streaming	Scala and Python	Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data
12	Impala	SQL	Cloudera's very own hive like interface but uses its own engines instead of relying on spark or map reduce engine for processing.

Now that we have the technology to skill mapping, lets translate this into a work schedule. I am assuming that you can spend 2 hours a day for 5 to 6 days a week. Given that Big Data is a mesmerizing world, i will not be surprised if you spend more than 2 hours a day purely out of interest and curiosity to learn. Hence a 6 week preparation should be good enough to crack the certification. I personally know people who did this in 2 weeks and hence nothing is impossible. So, there are just 6 weeks between the current "you" (possibly a no one in the big data context) to being someone i.e certified hadoop and spark developer. **Are you up to the challenge?** If the technologist and the curious learner inside you is urging you to shout **'YES I AM UP TO THE CHALLENGE'** then i recommend that you either spend the next few weeks in equipping yourself in these technologies and come back to this blog when trying to accomplish the 15th task in the schedule below OR use the videos in this blog to gain some understanding in a more real time learning environment where you learn concepts on the fly and in a hands-on fashion.
This series of blog posts will provide a series of mock problem exercises you can solve to test your skill and knowledge required for clearing the certification exam. Please explore the links in the menu (right side) and also go through the videos in the playlist.

#	Task	Hrs of Study	Hrs of Practice
1	Setup Cloudera quickstart VM	3 Hours	NA
2	Introduction to Hadoop and Basic understanding of what Big Data is in general	3 Hours	NA
3	HDFS	1 Hour	2 Hours
4	Sqoop	2 Hours	2 Hours
5	Scala	3 Hours	3 Hours
6	Python	3 Hours	3 Hours
7	Spark RDD	3 Hours	6 Hours
8	Spark DF	1 Hours	2 Hours
9	Spark SQL	1 Hours	2 Hours
10	Spark Submit	1 Hours	1 Hour
11	Flume	1 Hours	3 Hours
12	Spark Streaming	1 Hours	1 Hour
13	Hive	3 Hours	5 Hours
14	Impala	1 Hours	2 Hours
15	Scenarios	NA	4 Hours
16	Total	27 Hours	39 Hours
17	Grant Total	66 Hours	
18	Total Weeks of Prep at 2 Hours a day and 5 days a week	Around 6 Weeks	

GOOD LUCK.....

G+

49 comments:

SV May 8, 2017 at 1:04 PM

Thx for the helpful posts and the scenarios. Hopefully you will keep adding more content so users can clear the exam successfully!