**Agenda**

* CCA 175 – Introduction
* Learning Objectives
* Spark Introduction
* Preparation plan
* Resources

**CCA 175 Introduction**

* Certification is conducted by Cloudera
* It requires skills with respect to Sqoop, Spark, SQL/Hive
* Scenario-based
* Sqoop is command-oriented and one need to have good  
  understanding about sqoop commands
* For Spark, Programming skills are required – Either Python or Scala
* SQL skill is required
* It is not mandatory to provide the solution with one approach over  
  other.
* Cloudera does not see your code, they only evaluate results

**Learning Objectives**

* URL – [https://www.cloudera.com/more/training/certification/cca-spark.html 46](https://www.cloudera.com/more/training/certification/cca-spark.html)
* Data Ingest
* Transform, Stage and Store
* Data Analysis
* Configuration

**Spark Introduction**

* Spark is nothing but distributed processing engine
* It provides a bunch of APIs to facilitate distributed computing
* We need to use a programming language such as Scala or Python to  
  crack CCA 175 certification
* Spark also has high-level modules (eg: Spark SQL and Data  
  Frames, MLLib etc)
* For the certification, one should be able to understand Spark core  
  API as well as Spark SQL and Data Frames with some basic  
  understanding of Spark Streaming
* Majority of the questions from CCA 175 can be answered using  
  Spark and programming languages such as Scala or Python

**Preparation Plan**

* Understand basic HDFS commands
* Learn how to move data between relational databases and HDFS  
  using Sqoop
* Choose a programming language (Python or Scala)
  + Be comfortable with functions, lambda functions
  + Collections
  + Data Frames (Pandas in Python)
* Refresh SQL skills (preferably using Hive)
* Develop Spark based applications using Core APIs
  + Actions
  + Transformations
* Integrate Spark SQL and Data Frames to Spark based applications
* Make sure you understand Spark Streaming in tandem with Flume  
  and Kafka

**Resources**

* Cloudera Quickstart VM
  + Free
  + Requires high-end laptop (16 GB RAM, Quad-Core)
  + Might run into issues with respect to limited resources
* Big Data labs from itversity – [https://labs.itversity.com 22](https://labs.itversity.com/)
  + Support using [http://discuss.itversity.com 5](http://discuss.itversity.com/)
  + Multi-node cluster
  + Ability to access from anywhere
  + Pre-built datasets
  + Simulates the certification environment