Integrated Program In

BIG DATA and DATA SCIENCE

Duke | CONTINUING STUDIES

Table of Contents

About the Course

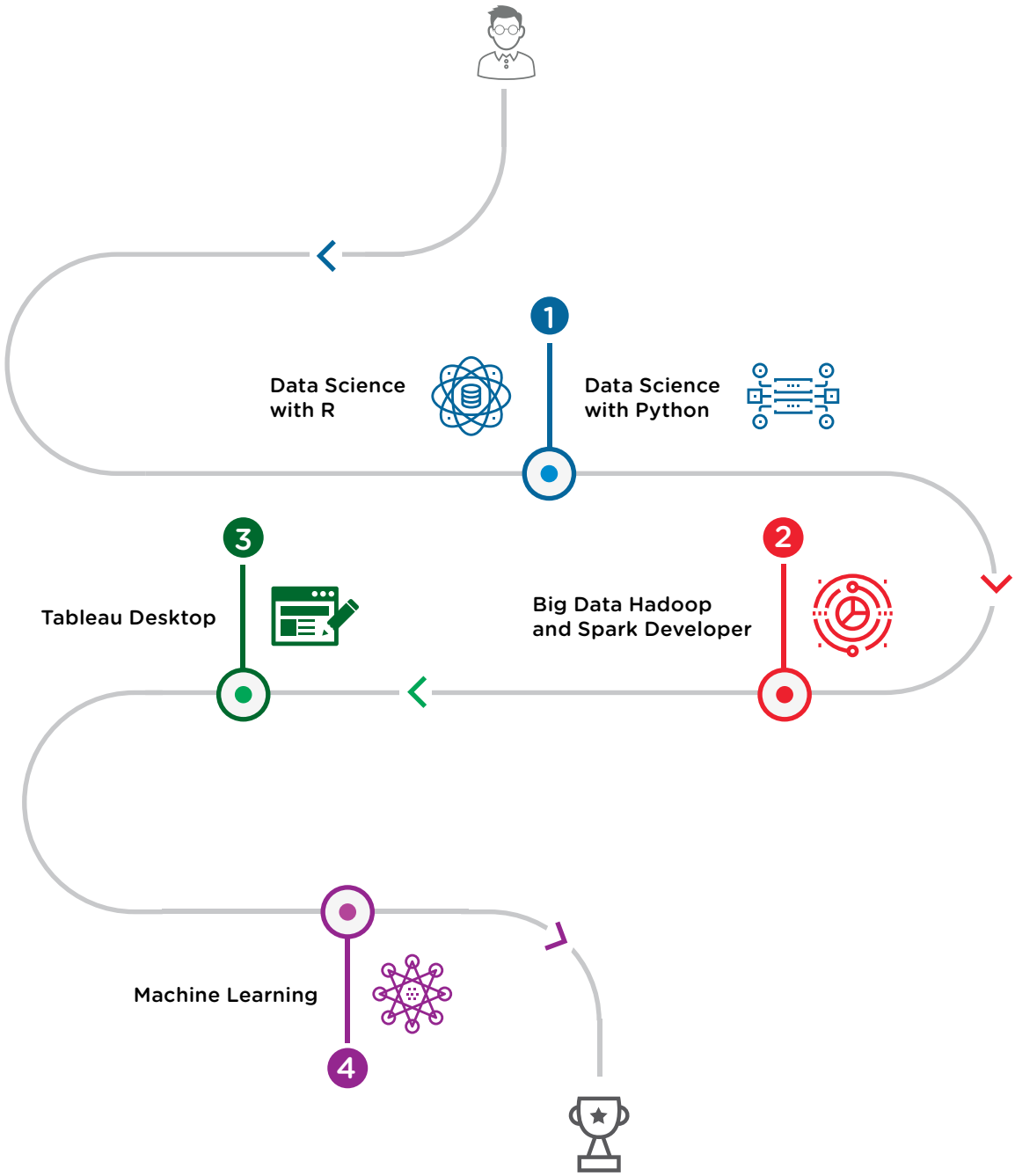
The Big Data and Data Science program is a five-course, integrated, all-inclusive certificate program for Big Data and Data Science professionals. The curriculum is comprehensive and spans the major technologies in big data, data science, and reporting/visualization. The recommended learning path for this certificate program has been designed by renowned industry experts and big data influences to maximize your learning potential. As each course of the program builds upon the next, concepts introduced initially in the learning path will contribute to your proficiency with concepts for the later courses of the program.

Resources such as live virtual teaching sessions, access to an instructor, and non-graded electives of your choosing reinforce this programs learning experience.

Key Features

- Industry-recommended learning path
- Access to 300+ hours of content created by industry experts
- Hands-on project execution on CloudLabs
- Aligned with the Cloudera CCA175 Certification and Tableau Desktop 10 Associate Certification
- Duke University Certificate upon successful completion of the course
- 30+ real-life industry projects in retail, insurance, healthcare, banking, telecommunication, airline and social media

Learning Path



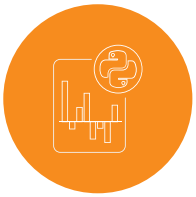
BIG DATA AND DATA SCIENCE

Key Learning Objectives

This learning path is designed for a professional interested in the field of analytics who wishes to develop skills in both big data and data science.

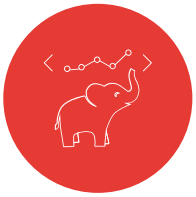
Data Science with R

Learn R programming language and all the important statistical and predictive analytics concepts



Data Science with Python

Introduces the various packages in Python like NumPy, SciPy, Pandas, and Scikit-learn for performing data analysis.



Big Data Hadoop and Spark Developer

Learn the various components of Hadoop and Spark ecosystem. The course is aligned to Cloudera CCA175 certification.

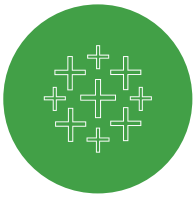


Tableau Desktop and Visualization Training

Learn the various aspects of Tableau. Aligned with Tableau Desktop Qualified Associate certification.

Machine Learning

Gain an understanding of Machine Learning applications and algorithms. It also covers deep learning and Spark Machine learning.

STEP 1 2 3 4

Data Science with R

This course has been designed to impart an in-depth knowledge of the various data analytics techniques that can be performed using R. It includes real-life projects, case studies, and R CloudLabs for practice.

Key Learning Objectives

- Gain a foundational understanding of business analytics.
- Learn the R programming and how various statements are executed.
- Gain an in-depth understanding of data structure used in R and learn to import/export data in R.
- Define and use the various apply functions and DPLYP functions.
- Recognize and use the various graphics in R for data visualization.
- Gain a basic understanding of the various statistical concepts.
- Understand the hypothesis testing method to drive business decisions.
- Become familiar with regression models and classification techniques.
- Learn and use the various association rules and the Apriori algorithm.
- Gain an understanding of clustering methods including K-means, DBSCAN, and hierarchical clustering.

STEP 1 2 3 4

Data Science with Python

Learn data analytics, machine learning, and web scraping using Python programming. Gain an in-depth understanding of the various packages in Python like NumPy, SciPy, Pandas, and Scikit-learn for performing data analysis, implementing machine learning models, and NLP.

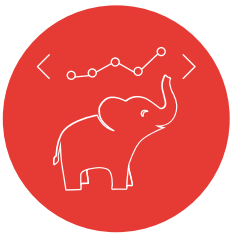
Key Learning Objectives

- Gain an in-depth understanding of data wrangling, data exploration, data visualization, hypothesis building, and testing.
- Understand the essential concepts of Python programming like data types, tuples, lists, dicts, basic operators, and functions.
- Perform high-level mathematical computing using NumPy package and its large library of mathematical functions.
- Conduct scientific and technical computing using SciPy package and its sub-packages such as Integrate, Optimize, Statistics, IO, and Weave.
- Perform data analysis and manipulation using data structures and tools provided in Pandas package.
- Gain knowledge in machine learning using the Scikit-Learn package.
- Use matplotlib library of Python for data visualization.
- Extract useful data from websites by performing web scrapping.
- Integrate Python with Hadoop, Spark, and MapReduce.

STEP 1 2 3 4

Big Data Hadoop & Spark Developer

This course has been designed to impart an in-depth knowledge of Big Data processing using Hadoop and Spark. The course contains real-life projects and case studies to be executed in CloudLabs and aligns with the Cloudera CCA175 certification.



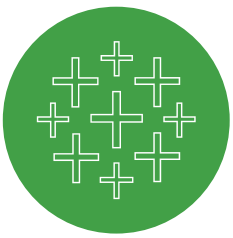
Key Learning Objectives

- Understand the architecture of HDFS and YARN, and learn how to work with them for storage and resource management.
- Recognize MapReduce and its characteristics.
- Receive an overview of Sqoop and Flume and how to ingest data.
- Create databases and tables in Hive and Impala, understand HBase, and use Hive and Impala for partitioning.
- Learn Flume architecture, sources, sinks and configurations.
- Understand HBase, its architecture, data storage.
- Gain a working knowledge of Pig and its components.
- Perform functional programming in Spark, understand RDDs and build Spark applications.
- Learn Spark SQL, and learn about creating, transforming, and querying data frames.
- Course aligned with the Cloudera Big Data CCA175 certification.

STEP 1 2 3 4

Tableau Desktop 10

The focus of the course is to help you learn Tableau Desktop 10 skills such as visualization building, analytics, and dashboards. This course is also aligned with the Tableau Desktop 10 Qualified Associate exam.



Key Learning Objectives

- Grasp the concepts of Tableau Desktop 10 and learn Tableau statistics and building interactive dashboards.
- Learn data connections as well as organizing and simplifying data.
- Understand formatting, annotations, and spatial analysis.
- Become familiar with special field types and Tableau generated fields.
- Review the concepts of using charts including Pareto, waterfall, Gantt, box plots, Sparkline and perform market basket analysis.
- Learn fundamental calculations along with automatic and custom split, ad-hoc analytics, and LOD calculations.
- Understand process of creating and using parameters and gain command over mapping concepts such as custom geocoding and radial selections.

STEP 1 2 3 4

Machine learning

This course provides advanced-level training on Machine Learning applications and algorithms. It will give you hands-on experience in multiple, highly sought-after machine learning skills in both supervised and unsupervised learning. This machine learning training helps you learn to apply machine learning algorithms like regression, clustering, classification, and recommendation. The unique case study approach ensures you are working hands-on with data while you learn. You'll also receive training in deep learning and Spark Machine learning—skills which are in high demand today.

Key Learning Objectives

- Classify the types of learning including supervised and unsupervised.
- Identify the various applications of machine learning algorithms.
- Perform supervised learning techniques: linear and logistic regression.
- Understand classification data and models.
- Use unsupervised learning algorithms including deep learning, clustering, and recommendation systems.
- Experience using machine learning with Spark.

Elective Courses

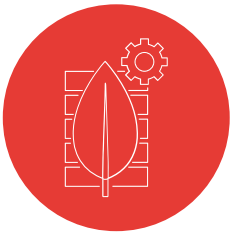
Data Science with SAS

The data science with SAS training is designed to impart an in-depth knowledge of SAS programming language, SAS tools, and various advanced analytics techniques.

Apache Spark and Scala

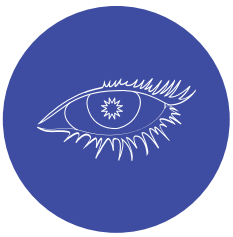
With this Apache Spark you will learn the essential skills such as Spark Streaming, Spark SQL, Machine Learning Programming, GraphX Programming, Shell Scripting Spark.

MongoDB Developer and Administrator



MongoDB training helps you learn data modelling, ingestion, query and Sharding, Data Replication with MongoDB along with installing, updating, and maintaining MongoDB environment.

Cassandra



The Apache Cassandra training provides you with in depthknowledge of Cassandra architecture, features, configuration and hadoop ecosystem around this NoSQL database.

Business Analytics with Excel

Business Analytics with Excel training has been designed to help initiate you to the world of analytics. For this we use the most commonly used analytics tool—Microsoft Excel. The training will equip you with the concepts and hard skills required to work in this industry.

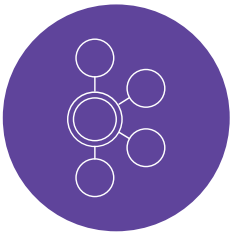
Apache Storm

Apache Storm training provides you with experience in stream processing Big Data technology of Apache Storm.

Impala: An Open Source SQL Engine for Hadoop Training Course

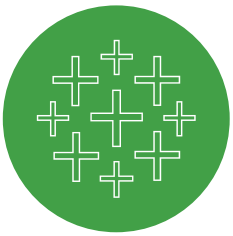
The “Impala: An Open Source SQL Engine for Hadoop” is an ideal course package for individuals who want to understand the basic concepts of Massively Parallel Processing or MPP SQL query engine that runs on Apache Hadoop. Upon completing this course, learners will be able to interpret the role of Impala in the Big Data Ecosystem.

Apache Kafka



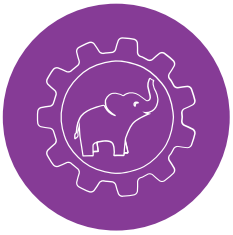
The Apache Kafka course guides participants through the Kafka architecture, installation, interfaces, and configuration. The participants are also trained in the fundamental concepts of Big Data in this course.

Tableau Server 10 Qualified Associate



The Tableau Server 10 Qualified Associate course is designed to impart in-depth understanding and skills to implement, administer, and manage Tableau 10 server. This course is designed for Tableau server users and administrators.

Big Data Hadoop Administrator



Big Data and Hadoop Administrator course is aligned with Cloudera's CCAH "CCA-500" certification and covers the core Hadoop distributions—Apache Hadoop and Vendor specific distribution—CDH (Cloudera Distribution of Hadoop).

Instructors

Ronald Van Loon

Top 10 Big Data & Data Science Influencer,
Director - Adversitement

Named by Onalytica as one of the three most influential people in Big Data, Ronald writes for a number of leading Big Data and Data Science websites, including Datafloq, Data Science Central, and The Guardian. He is a regular speaker at renowned events.

Sina Jamshidi

Big Data Lead at Bell Labs

Sina has over 10 years of experience in the Technology field as a Big Data Architect at Bell Labs and as a Platinum-level trainer. Sina is a very passionate about building a Big Data education ecosystem and has been a contributor in a number of public and journal publications.

Simon Tavasoli

Analytics Lead at Cancer Care, Ontario

Simon is a Data Scientist with 12 years of experience in Healthcare Analytics. He has a Masters in Biostatistics from the University of Western Ontario. Simon is passionate about teaching data science and has published several journals in preventive medicine analytics.

Instructors

Alvaro Fuentes

Founder and Data Scientist at Quant Company

Alvaro is a Data Scientist who founded Quant Company and has also worked as a lead Economic analyst in the Central Bank of Guatemala. He is a M.S. in Quantitative Economics and Applied Mathematics and is actively involved in consulting and training in the data science space.

Paul Sharkov

Data Scientist at BMO Financial Group, Member of SAS Canada Community

Paul is lead SAS Data Scientist at Bank of Montreal. As a SAS Certified Predictive Modeler, SAS Statistical Business Analyst, and SAS Certified Advanced Programmer, Paul is passionate about sharing his knowledge on how data science can support data-driven business decisions.

Live virtual classrooms are facilitated by qualified industry subject matter experts in alignment with the curriculum designed by the instructors listed above.

Duke | CONTINUING STUDIES

Duke Continuing Studies

Box 90700
Duke University East Campus
Durham, NC 27708-0700
(919) 684-6259
learnmore@duke.edu
http://learnmore.duke.edu