



## SAS Certified Data Scientist



The SAS Certified Data Scientist program – offered by the SAS Academy for Data Science – can deepen your knowledge, jump-start your career and boost your earning power. SAS Certified Data Scientist Program is divided into 2 levels:

**Level 1: Big Data Professional**

**Level 2: Advanced Analytics Professional**

### ➤ Curriculum

The data science certification program curriculum includes the following courses:

#### Overview of Courses of Level 1: Big Data Professional

##### **Big Data Challenges and Analysis-Driven Data**

This course provides an overview of the challenges associated with big data and analysis-driven data.

###### *Topics Covered*

- Reading external data files.
- Storing and processing data.
- Combining Hadoop and SAS.
- Recognizing and overcoming big data challenges.

##### **Exploring Data with SAS Visual Analytics**

In this course, you'll learn how to use SAS Visual Analytics Explorer to explore in-memory tables from the SAS® LASR™ Analytic Server and perform advanced data analyses.

###### *Topics Covered*

- Finding previously unknown relationships and spotting trends in your data.
- Visualizing data using charts, plots and tables.
- Using the autocharting function to visualize data in the best possible way.
- Using advanced graphs, such as network diagrams, Sankey diagrams and word clouds.
- Easily adding analytics to your graphs, and including descriptions of the analytics results.
- Navigating through your data using on-the-fly hierarchies.

##### **Preparing Data for Analysis and Reporting**

In this course, you'll learn how to perform data management tasks, such as improving data quality, entity resolution and data monitoring.

###### *Topics Covered*

- Creating and reviewing data explorations.
- Creating and reviewing data profiles.
- Creating data jobs for data improvement.
- Establishing monitoring aspects for your data.
- Understanding the QKB components.
- Using the component editors.
- Understanding various definition types.
- Building a new data type (optional).



## SAS Certified Data Scientist



### Statistics 1: Introduction to ANOVA, Regression and Logistic Regression

This introductory SAS/STAT® course focuses on t-tests, ANOVA and linear regression, and includes a brief introduction to logistic regression.

#### Topics Covered

- Generating descriptive statistics and exploring data with graphs.
- Performing analysis of variance and applying multiple comparison techniques.
- Performing linear regression and assessing the assumptions.
- Using regression model selection techniques to aid in the choice of predictor variables in multiple regression.
- Using diagnostic statistics to assess statistical assumptions and identify potential outliers in multiple regression.
- Using chi-square statistics to detect associations among categorical variables.
- Fitting a multiple logistic regression model.
- Scoring new data using developed models.

### Introduction to SAS and Hadoop: Essentials

This course teaches you how to use SAS programming methods to read, write and manipulate Hadoop data. You'll learn how to use Base SAS methods to read and write raw data with the DATA step, manage the Hadoop Distributed File System (HDFS) and execute MapReduce and Pig code from SAS via the HADOOP procedure. You'll also learn how to use SAS/ACCESS® Interface to Hadoop methods that allow LIBNAME access and SQL pass-through techniques to read and write Hive or Impala table structures. *Topics Covered*

- Accessing Hadoop distributions using the LIBNAME statement and the SQL pass-through facility.
- Creating and using SQL procedure pass-through queries.
- Using options and efficiency techniques for optimizing data access performance.
- Joining data using the SQL procedure and the DATA step.
- Reading and writing Hadoop files with the FILENAME statement.
- Executing and using Hadoop commands with PROC HADOOP.
- Using Base SAS procedures with Hadoop.

### DS2 Programming Essentials with Hadoop

This course focuses on DS2, a fourth-generation SAS proprietary language for advanced data manipulation, which enables parallel processing and storage of large data with reusable methods and packages.

#### Topics Covered

- Identifying the similarities and differences between the SAS DATA step and the DS2 DATA step.
- Converting a Base SAS DATA step to DS2.
- Creating DS2 variable declarations, expressions and methods for data conversion, manipulation and conditional processing.
- Creating user-defined and predefined packages to store, share and execute DS2 methods.
- Creating and executing DS2 threads for parallel processing.
- Leveraging the SAS In-Database Code Accelerator to execute DS2 code outside of a SAS session.
- Executing DS2 code in the SAS High-Performance Analytics grid using the HPDS2 procedure.



### Big Data Analysis with Hive and Pig

In this hands-on course, you'll use processing and analysis to find insights in structured and unstructured big data. You'll learn how to organize structural data in tabular format using Apache Hive and how to analyze the data using the Hive query language (HiveQL). You'll use the Apache Pig scripting language to perform batch processing tasks, such as extract, transform, load (ETL), data preparation and analytics.

#### Topics Covered

- Moving data into the Hadoop ecosystem.
- Using Hive to design a data warehouse in Hadoop.
- Performing data analysis using HiveQL.
- Joining data sources.
- Performing ETL.
- Organizing data in Hadoop by usage.
- Performing analysis on unstructured data using Pig.
- Joining massive data sets using Pig.
- Using user-defined functions (UDFs).
- Analyzing big data in Hadoop using Hive and Pig.

### Getting Started with SAS In-Memory Statistics

This course focuses on accessing data on the SAS LASR Analytic Server and performing exploratory analysis and preparation. Topics include starting the server, loading data and manipulating data on the SAS LASR Analytic Server using the IMSTAT procedure. IMSTAT topics include deriving new temporary and permanent tables and columns, calculating summary statistics (e.g., mean, frequency and percentile), and creating filters and joins on in-memory data.

#### Topics Covered

- Starting up a SAS LASR Analytic Server.
- Loading tables into memory on the SAS LASR Analytic Server.
- Processing in-memory tables with PROC LASR and PROC IMSTAT.
- Accessing data more efficiently via intelligent partitioning.
- Deriving new temporary and permanent tables and variables.
- Creating filters and joins on in-memory data.
- Exporting ODS result tables for client-side graphic development.
- Producing descriptive statistics including counts, percentiles and means.
- Creating multidimensional summaries including cross-tabulations and contingency tables.
- Deriving kernel density estimates using normal functions.

### Review and Case Study Level 1

- C1 Review: Programming Syntax and Case Study 1
- C2 Review: Exploring Data with SAS Visual Analytics and Case Study 2
- C3 Review: Statistics 1: Introduction to ANOVA, Regression and Logistic Regression and Case Study 3
- C4 Review : Preparing Data for Analysis and Reporting and Case Study 4
- C5 Review: Introduction to SAS and Hadoop and DS2 Essentials with Hadoop and Case Study 5
- C6 Review : Big Data Analysis with Hive and Pig and Case Study 6
- C7 Review: Getting Started with SAS In-Memory Statistics and Case Study 7



### Overview of Courses of Level 2: Advanced Analytics Professional

#### Applied Analytics Using SAS Enterprise Miner

This course covers the skills required to assemble analysis flow diagrams using SAS Enterprise Miner for both pattern discovery (segmentation, association and sequence analyses) and predictive modeling (decision trees, regression and neural network models).

##### Topics Covered

- Defining a SAS Enterprise Miner project and exploring data graphically.
- Modifying data for better analysis results.
- Building and understanding predictive models, including decision trees and regression models.
- Comparing and explaining complex models.
- Generating and using score code.
- Applying association and sequence discovery to transaction data.

#### Communicating Technical Findings to a Non-Technical Audience

This course teaches you how to design and communicate effective presentations through self-assessment and discussions about presentation organization and the effective use of visual aids. You will receive an individual analysis of your behavioral style, including a description of your strengths and opportunities for improvement, as well as strategies for communicating with others.

##### Topics Covered

- Diagnosing and assessing different styles of human behavior.
- Communicating and coping more effectively with different types of people.
- Using your own strengths and knowledge of others to enhance communication.
- Delivering information in a concise and well-organized format.
- Creating a presentation, with the focus on communicating unfamiliar or technical information to a nontechnical audience.
- Designing presentation materials with clarity and purpose.

#### Neural Network Modeling

This course helps you understand and apply two popular artificial neural network algorithms – multilayer perceptrons and radial basis functions. Both the theoretical and practical issues of fitting neural networks are covered.

##### Topics Covered

- Constructing multilayer perceptron and radial basis function neural networks.
- Constructing custom neural networks using the NEURAL procedure.
- Choosing an appropriate network architecture and determining the relevant training method.
- Avoiding overfitting neural networks.
- Performing autoregressive time series analysis using neural networks.
- Interpreting neural network models

#### Predictive Modeling Using Logistic Regression

This course explores predictive modeling using SAS/STAT® software, with an emphasis on the LOGISTIC procedure.

##### Topics Covered



## SAS Certified Data Scientist



- Using logistic regression to model an individual's behavior as a function of known inputs.
- Selecting variables and interactions.
- Creating effect plots and odds ratio plots using ODS Statistical Graphics.
- Handling missing data values.
- Tackling multicollinearity in your predictors.
- Assessing model performance and comparing models.
- Recoding categorical variables based on the smooth weight of evidence.
- Using efficiency techniques for massive data sets.

### Data Mining Techniques: Predictive Analytics on Big Data

This course introduces applications and techniques for assaying and modeling large data. It presents basic and advanced modeling strategies, such as group-by processing for linear models, random forests, generalized linear models and mixture distribution models. You will perform hands-on exploration and analyses using tools such as SAS Enterprise Miner, SAS Visual Statistics and SAS In-Memory Statistics.

#### Topics Covered

- Using applications designed for big data analyses.
- Exploring data efficiently.
- Reducing data dimensionality.
- Building predictive models using decision trees, regressions, generalized linear models, random forests and support vector machines.
- Building models that handle multiple targets.
- Assessing model performance.
- Implementing models and scoring new predictions.

### Using SAS to Put Open Source Models into Production

This course introduces the basics for integrating R programming and Python scripts into SAS and SAS Enterprise Miner. Topics are presented in the context of data mining, which includes data exploration, model prototyping, and supervised and unsupervised learning techniques.

#### Topics Covered

- Calling R packages in SAS.
- Leveraging Python scripts in SAS.
- Integrating open source data exploration techniques in SAS.
- Integrating open source models in SAS Enterprise Miner.
- Creating production (score) code for R models.

### Text Analytics Using SAS Text Miner

In this course, you will learn to use SAS Text Miner to uncover underlying themes or concepts contained in large document collections, automatically group documents into topical clusters, classify documents into predefined categories, and integrate text data with structured data to enrich predictive modeling endeavors.

#### Topics Covered

- Converting documents stored in standard formats (Microsoft Word, Adobe PDF, etc.) into general-purpose HTML or TXT formats.
- Reading documents from a variety of sources (web pages, flat files, data elements in a relational database, spreadsheet cells, etc.) into SAS tables.
- Processing textual data for text mining (e.g., correcting misspellings or recoding acronyms and abbreviations).



## SAS Certified Data Scientist



- Converting unstructured text-based character data into structured numeric data.
- Exploring words and phrases in a document collection.
- Querying document collections using keywords (i.e., identifying documents that include specific words or phrases).
- Identifying topics or concepts that appear in a document collection.
- Creating user-influenced topic tables from scratch or by modifying machine-generated topics, or creating concepts using domain knowledge.
- Using derived topic tables or pre-existing user-influenced topic tables (or both) to enhance information retrieval and document classification.
- Clustering documents into homogeneous subgroups.
- Classifying documents into predefined categories.

### Time Series Modeling Essentials

In this course, you'll learn the fundamentals of modeling time series data, with a focus on the applied use of the three main model types for analyzing univariate time series: exponential smoothing, autoregressive integrated moving average with exogenous variables (ARIMAX), and unobserved components (UCM).

#### Topics Covered

- Creating time series data.
- Accommodating trend, as well as seasonal and event-related variation, in time series models.
- Diagnosing, fitting and interpreting exponential smoothing, ARIMAX and UCM models.
- Identifying relative strengths and weaknesses of the three model types.

### Experimentation in Data Science

This course explores the essentials of experimentation in data science, why experiments are central to any data science efforts, and how to design efficient and effective experiments.

#### Topics Covered

- Defining common terminology in designed experiments.
- Describing the benefits of multifactor experiments.
- Differentiating between the impact of a model and the impact of the action taken from that model.
- Fitting incremental response models to evaluate the unique contribution of a marketing message, action, intervention or process change on outcomes.

### Optimization Concepts for Data Science

This course focuses on linear, nonlinear and efficiency optimization concepts. Participants will learn how to formulate optimization problems and how to make their formulations efficient by using index sets and arrays. Course demonstrations include examples of data envelopment analysis and portfolio optimization. The OPTMODEL procedure is used to solve optimization problems that reinforce concepts introduced in the course.

#### Topics Covered

- Identifying and formulating appropriate approaches to solving various linear and nonlinear optimization problems.
- Creating optimization models commonly used in industry.
- Formulating and solving a data envelopment analysis.
- Solving optimization problems using the OPTMODEL procedure in SAS.



## SAS Certified Data Scientist



### Review and Case Study Level 2

- C1 Review: Applied Analytics Using SAS Enterprise Miner and Case Study 1
- C2 Review: Neural Network Modeling and Case Study 2
- C3 Review: Predictive Modeling Using Logistic Regression

### ➤ Certification exams

The data science certification program includes five certification exams. To earn the SAS Certified Data Scientist credential, you must pass all five exams:

- [SAS Big Data Preparation, Statistics and Visual Exploration](#)
- [SAS Big Data Programming and Loading](#)
- [Predictive Modeling Using SAS Enterprise Miner 13\\*](#)
- [SAS Advanced Predictive Modeling](#)
- [SAS Text Analytics, Time Series, Experimentation and Optimization](#)

### Key Value Propositions for Individuals-

