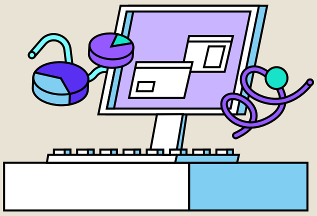
DataScientest

SYLLABUS DATA ENGINEER



DISCOVER THE DIFFERENT JOBS IN DATA

OUR STORY

DataScientest is the leading training organization in Europe for Data Science. We offer business-oriented training courses for upskilling and reskilling to professionals and individuals alike.

TRAINING DESIGNED WITH AND FOR COMPANIES

Data has become the main resource to exploit for companies. It allows them to **guide their decision-making** and to predict the behaviour of their environment. Therefore, staying **competitive** means exploiting data to its fullest potential. Thus, given that data is everywhere, it is essential to **increase workforce capabilities** on these new skills, tools and technologies to ensure stable development.

This is why DataScientest has been training the Data teams of more than **70 Fortune 500 groups** as well as many other SMEs and Startups for over **6 years**. Since 2021, more than **5500 professionals** have been trained with us.

74%

of the heads of Data in the largest groups wish to strengthen their teams by recruiting Data Engineers*

*Survey conducted among our forty partner companies.

A CURRICULUM ADAPTED TO INDIVIDUALS

Having been designed for our **partner companies**, our training courses are professionally oriented, as close as possible to the **needs of the market**. Thus, in order to face the shortage of Data profiles and bridge this gap, we have decided to open our training courses to **individuals**. Under this impulse, DataScientest is constantly developing thanks to **new training courses** (Data Manager) or **new specialized courses** (Deep Learning).

> *"DataScientest is a rich, personalized learning experience that responds to the real problems of the company.*
>
> Samir Ait Idir
> Chief Data Officer & Head of Data Intelligence @Orange Bank

DATASCIENTEST KEY FIGURES

+70 **FORTUNE 500 COMPANIES** AS CLIENTS

98% **COMPLETION** RATE

94% **SATISFACTION** RATE

+5500 **ALUMNI**

2000h **OF EXCLUSIVE** CONTENT

THEY TRUST US

GROUPE ADP

STELLANTIS

arianeGroup

AXA

GROUPE BPCE

BCG

BNP PARIBAS CARDIF

bouygues TELECOM

Capgemini

CA

eDF

ENEDIS L'ELECTRICITE EN RESEAU

MICHELIN

orange™

SAFRAN

TotalEnergies

OUR PARTNERS

DataScientest has developed partnerships with academic institutions on the one hand and with software publishers on the other. The content of our training courses on Business Intelligence or Cloud tools is henceforth certified by the companies that own them. Here are our different partnerships:

AMAZON WEB SERVICES
Technological Partner

Today, DataScientest enjoys the exclusive status of **Amazon Training Partner**. We are therefore authorized by Amazon to train teams on the products and services of the American company. Thanks to this partnership, we have established **several training courses** that prepare for **official AWS certifications**. The registration fee for the official exam is included in the price of the course.
In some of our courses, you will be prepared to take the fundamental **AWS certification: Cloud Practitioner.**

MICROSOFT - POWER BI
Technological Partner

DataScientest is a **Microsoft Learning Partner**, which means that we can train you for **official Microsoft certifications**. These certifications attest to a level of expertise in Azure, the set of cloud computing products and services, and in **Power BI**, Microsoft's business intelligence tool. The registration fee for the official exam is included in the price of the course.
In each training course, DataScientest prepares you to pass official Microsoft certifications. Additionally, with the **Data Engineer** course, DataScientest allows you to prepare for the **A9-900** certification, **Microsoft Certified Azure Fundamentals**.

L'ÉCOLE DES MINES - PARISTECH EXED
Academic Partner

Ecole des **Mines - PSL Executive Education** is a **top-ranked engineering school** with one of the most dynamic Data Science academies in Europe. DataScientest is proud to be an **official partner**, which attests to the **quality of our Data Scientist and Data Engineer programs**, which are eligible for **official certification by the school**.

QUALIOPI
Quality Criteria

This label recognises that the processes implemented by the training organization meet the quality requirements demanded by the State. **All DataScientest training courses are Qualiopi certified.**

OUR TEAM

OUR TEACHERS

DataScientest's teaching team is made up of internal professors. They are dedicated to teaching and research for our various courses and expert curricula.

All our professors are always there for you. They have created and continue to update all our courses. They will accompany you by video-conference and on our platform throughout your training.
With their excellent academic background and varied professional experience, they are the Data Science experts who will enable you to join the Data team in the sector that interests you most (Banking & Insurance, Industry, Finance, Medical and many others).

Charles S.
Head of Academic
8 years of experience

A graduate of the École Polytechnique, Charles has been involved in the development of the courses from day one. He is specialized in programming, Machine Learning and Deep Learning.

Raphael K.
Pedagogical Director
8 years of experience

Holder of the ISF master's degree specializing in Statistical Learning and Data Science from the University of Paris-Dauphine, Raphaël designed the Data Analyst and Data Management courses thanks to his knowledge in programming, dataviz and Machine Learning.

MEET OUR TEAM

OUR TEAM

Frédéric F.
Data Engineer course referent
2 years of experience

Frédéric holds a master's degree from the University of Paris-Dauphine PSL Executive. A few years ago, he joined the DataScientest team and specialized in big data. He is now in charge of our Data Engineer training and of the expert course Approfondissement Engineering.

Thomas B.
Deep Learning course director
6 years of experience

After attending a top tier engineering school, Thomas quickly entered DataScientest's ranks. His array of knowledge stretches from programming and Dataviz to NLP, Deep Learning and Computer vision.

DANIEL

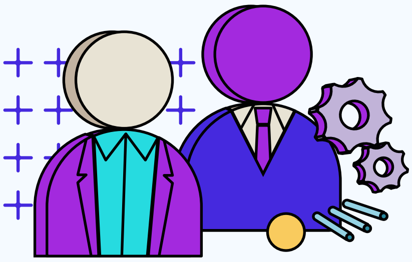
During your training, you will progress on our platform. However, **you will never be alone**, Daniel is there for you. All our teachers and Program Managers take turns to **answer all your questions live via Slack** and on a dedicated support platform. With support every day of the week, from 9am to 5pm, you are guided throughout your course.

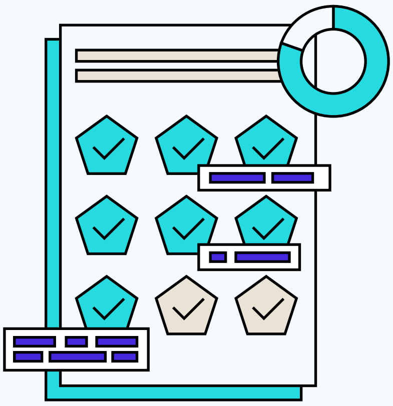
OUR PEDAGOGY

HYBRID FORMAT

DataScientest offers 100% distance learning in a hybrid format :

During approximately 20% of the duration, you are accompanied by your teacher and your cohort during **coaching sessions via videoconference**. The rest of the time, you work on our teaching platform and are **guided** via Slack by our **Data Scientists**. This format gives you the flexibility to organize yourself. We also monitor your progress and accompany you to ensure the **successful completion** of your training.





MASTERCLASS

Each training period on the platform (called a Sprint) is accompanied by one or more **masterclasses by videoconference**. The objectives and teaching formats of each masterclass change according to the sprint: correction of concrete use cases, lectures on specific topics, competitions between cohorts, etc.

EXAMENS

At DataScientest, there are no automated MCQs. **Each paper is reviewed and corrected** by hand by our teachers. They go over all your difficulties with you.

TRAINING LEADING TO CERTIFICATION - STATE RECOGNITION

The validation of the skills developed during our Data Engineer training will allow you to obtain a certificate from Mines ParisTech PSL Executive Education and to validate the «Deploy an artificial intelligence solution» **block of competences of the RNCP36129 certification recognized by the State as a Bac +5 (European level 7)**. This is a strong signal on the job market.

FRANCE compétences CERTIFICATION enregistrée au RNCP

BACKBONE PROJECT

From the beginning of your training, you will carry out a concrete project with the aim to deploy it. It requires an investment of about 120 hours of work throughout the training.

You will work in pairs or in triads on a project that you will select from our regularly updated project **catalog**. Our subjects are inspired by the work we do with companies.
This is a crucial step in your **career path** that will make you fully operational. You will be working on uncleaned data sets, and will be expected to produce work of professional quality.

Reviews are done at regular intervals by your teacher to guide you, and **coach you**. This allows you to move efficiently from **theory to practice** and to ensure that you master the **skills** required on the different modules.

It is also a project that is highly valued by companies. It confirms your skills and knowledge acquired at the end of the **Data Engineer** course. You will then be able to justify your skills with a **successful data science project** during your interviews.

> *DataScientest is a rich, personalized course, perfectly adapted to today's corporate data science topics.*
>
> Xavier Bocher
> Head of Credit Risk Internal Models & Operational Research @Groupe Crédit Agricole

PREREQUISITES

To be able to access the training program, it is necessary to have a diploma or RNCP title of Bac +3 in mathematics or Bac +5 in a scientific field. It is also required to demonstrate an understanding of SQL language and Linux systems.

In addition, this job is focused on data, it requires a solid foundation in mathematics and statistics, which will also allow you to grasp with greater ease new concepts that you will learn during the course.

CONTINUOUS FORMAT

Only one format is available for the Data Engineer training :

CONTINUOUS FORMAT

Flexible, it is the format adapted for those that work a full-time profession.

DURATION 9 months

RYTHM 10-12h / week

BOOK AN APPOINTMENT

THE CURRICULUM

Data Engineer

1 PROGRAMMING

NumPy
pandas

- Bash
- Python
- Object Oriented Python

2 DATA BASES

- SQL
- MangoDB
- ElasticSearch
- Neo4J

elasticsearch
neo4j

3 DATA SCIENCE

- Statistics
- Machine Learning
- Data Visualization

4 BIG DATA VOLUME

- Hadoop & Hive
- Hbase
- Spark

hadoop
HIVE
APACHE HBASE
Spark

5 CI/CD

git
GitHub

- Git
- GitHub
- Quality Assurance

6 BIG DATA VITESSE

- Streaming Architecture
- Kafka
- Spark Streaming

TensorFlow

7 AUTOMATION AND DEPLOYMENT

docker
Apache Airflow
kubernetes

- APIs
- Airflow
- Docker
- Kubernetes

LEARN MORE

PROGRAMMING - duration 40h

Linux system and Bash Script

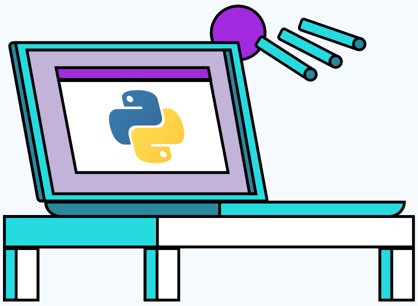
- Presentation of Linux Systems
- Handling and use of a terminal
- Setting up Bash scripts

Python & Object Oriented Python

- Mastering variables and types
- Presentation of the various operators and their and their applications
- Introduction to the concept of loops and control structures
- Definition of a function on Python and presentation of their applications
- Introduction to classes and modules
- Preparation of the implementation, the parameterization and the chaining of Decorators
- Differentiation and implementation of multithreading and multiprocessing on Python
- Application of an asynchronous function on Python
- Introduction to annotations and use of the of the MyPy library

Skills acquired at the end

- Master the Linux operating system
- Learn to use a Terminal
- Create and manage Bash executables
- Master the Python language and all its applications
- Understand and use object-oriented object-oriented programming
- Create complex scripts with Python



DATA BASES - duration 50h

SQL

- Introduction to relational databases relational databases
- Presentation of SQL Alchemy and applications
- Introduction to the basics of the SQL language
- Learning SQL and its applications

MongoDB

- Introduction to databases NoSQL (document, column, graph oriented databases document, column, graph oriented databases)
- Presentation of MongoDB
- Familiarization with the syntax of MongoDB queries

ElasticSearch

- Description of a search engine
- Presentation of an index and instructions for use
- Development of a Mapping
- Discovery of the different operations
- Data pre-processing with Ingest Node
- Extraction of data with the Text analyzer

Neo4J

- Introduction to graph-oriented databases oriented databases
- Setting up a first graph
- Introduction to the Cypher query language
- Loading data into Neo4J
- Using a Python client for Neo4J

Skills acquired at the end

- Know how to choose a database management system according to the use case
- Understand the notion of schemas and their implementation in a relational database
- Understand how to query a RDBMS (relational database management system) with the SQL language
- Handle a document-oriented database like MongoDB
- Improve your textual data search using Elasticsearch
- Manage a graph-oriented database
- Use Cypher to query and update graph-oriented databases

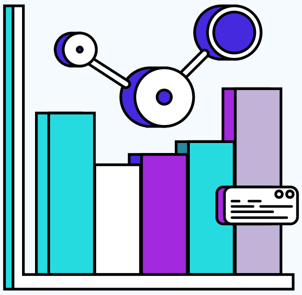
DATA SCIENCE - duration 50h

Statistics

- Exploration of numerical variables
- Exploration of categorical variables
- Study of relationships between variables

Machine Learning

- Data pre-processing
- Selection and optimization of a Machine Learning algorithm
- Definition and application of a regression regression algorithm
- Definition and application of a classification algorithm
- Development of clustering
- Introduction to PCA



Data Visualization with Matplotlib

- Presentation of different types of graphs:
  - Bar graphs (Barplots)
  - Scatter plots
  - Histograms
  - Box plots
  - Pie Plots
- Dash application creation

Skills acquired at the end

- Understand the basics of the main Machine Learning algorithms
- Be directly operational in machine learning
- Train machine learning models with the learning models with the Sckit-Learn library
- Manipulate your data with Pandas dataframes
- Mastering Numpy
- Visualize your data in various graphs with Matplotlib

BIG DATA VOLUME - duration 50h

Hadoop & Hive

- How Hadoop works
- Installation and configuration of Hadoop
- Data processing and storage with HDFS
- Introduction to MapReduce
- Using Hadoop Streaming to run a Map/ Reduce file
- Setting up data warehouses
- Presentation of how Hive works

Spark

- Distinction between Spark and Hadoop
- Introduction to distributed computing with Spark
- Overview of Spark's RDD and Dataframe APIs
- Distributed Data Processing Pipeline with PySpark
- Distributed Machine Learning with Spark MLLib

HBase

- Presentation of column-oriented databases
- Association of Hadoop (HDFS) and Hbase
- Data queries
- Data modification by Python and happybase

Skills acquired at the end

- Understand the fundamental concepts of Big Data
- Understand the theory of distributed systems architectures
- Store and process data in a distributed way data with Hadoop distributed file system (HDFS)
- Master the main tools for the management of management of Big Data:
  - Barplots
  - Scatter plots
  - Histograms
  - Box plots
  - Pie Plots

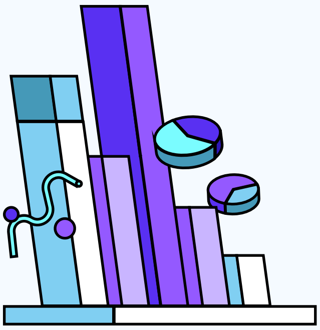
CI / CD - duration 15h

Git

- Introduction to the version management system
- Initialization of a Git repository
- Presentation and deepening of git concepts :
  - Branches
  - Tag
  - Merge

GitHub

- Implementation of unit tests with Pytest
- Introduction to Integration Tests and their functions
- Presentation of the advantages of testing time saving, readability, quality and quality and improvement of code



Quality Assurance

- Discover the Github platform for collaborative work on Git
- Presentation of the major features of GitHub :
  - Fork
  - Pull Request
  - Issues
- Share your modifications with pull and push
- Participation in the improvement of projects public (open source)
- Overview of main git workflows

Skills acquired at the end

- Mastering versioning tools
- Work collaboratively and version projects with Git and GitHub
- Be able to set up unit tests
- Apply methods adapted to the different issues
- Verify the functioning of independent code units during development

BIG DATA VITESSE - duration 15h

Streaming architecture

- Real-time data flow management
- Design of a hybrid Big Data architecture (batch and real time)
- Implementation of a Lambda architecture

Kafka

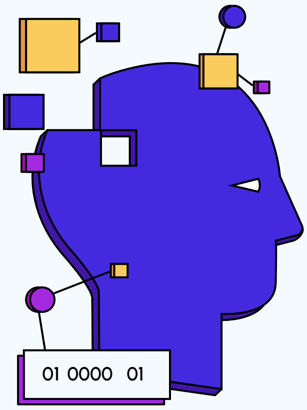
- Getting started with Spark Streaming for real-time data processing
- Presentation of the mini-batch streaming necessary for the operation of Spark Streaming

Spark Streaming

- Presentation of the distributed streaming platform Kafka :
  - Architecture
  - Advantages
- Management of Producers settings
  - Partitioning key
- Mastering Consumers settings
  - Consumer group

Skills acquired at the end

- Understand how to manage real-time data streams
- Implement and manage streaming architecture
- Process data in real time
- Master the Kafka software
- Process and transform real-time data in a distributed way with Spark Streaming



AUTOMATION AND DEPLOYMENT - duration 60h

APIs

- Introduction to APIs and discovery of microservices architectures
- Presentation of the different HTTP methods and their functions
- Use of the FastAPI and Flask libraries to develop RESTful APIs
- Documentation of an API with the OpenAPI specification
- Error and performance management of an API

Docker

- Presentation of containerization and its usefulness in relation to virtualization
- Introduction to the functioning of Docker
- Handling images and containers
- Communication with containers
- Data persistence thanks to volumes
- Creating a Docker image via a Dockerfile
- Sharing images on the Dockerhub
- Use of docker-compose

BOOK AN APPOINTMENT

Airflow

- Discovery of the Airflow concepts:
  - Presentation of the principles of orchestration principles and usefulness
  - Directed Acyclic Graphs or DAG (Directed Acyclic Graphs)
  - Operators
  - Task management through specific Operators
  - Monitoring of DAGs via the graphical graphical interface

Kubernetes

- Deploy and manage containers

Skills acquired at the end

- Understanding APIs
- Learn how to create an Api with Flask and FastApi
- Requesting an HTTP API
- Automate your tasks with Apache Airflow
- Understand virtualization
- Master the techniques and tools of containerization and container orchestration

ALUMNI TESTIMONIES

Steve NICOLE
Structural Engineer at Framatome

I did my training to become a Data Engineer at DataScientest. The training material is of high quality, the teaching team is invested, reactive and concerned about the success of its learners. Data engineer training is very demanding but rewarding, I recommend !

Hana MOUJOU

The training is rich, exciting and very demanding!!! A lot of pedagogy in the online courses and masterclasses as well as a real coherence in the pedagogical path. It follows Data Enginerie projects on various DevOps and Big Data technologies. The teaching staff is nice, available and reactive. In short, a training that I highly recommend!

Bachir ATMANI
Data Engineer at Mydral

I have taken the Data Engineer training and I find the two modules that deal with Machine Learning and Dataviz to be well done. The content is of good quality and it allows you to quickly become more competence !

Sofian Gaide
Data Engineer Intern at Ippon Technologies

3 months of intense but extremely rewarding training! See Data from A to Z in order to be ready to work with all kinds of data in business !
I highly recommend this training !

TO GO FURTHER

If you wish to strengthen your skills, DataScientest has a variety of training courses on publisher certifications such as Microsoft or AWS that enable you to deepen your knowledge and improve your skills in Data!

AZURE DATA ENGINEER

If you want to improve on the AZ-900 training and know how to implement solutions on Azure in an enterprise, this training is for you! At the end of the training, pass the official Microsoft certification and become an **«Azure Data Engineer Associate»**.

JE DÉCOUVRE LE CURSUS

AZURE ADMINISTRATOR

If you want to have skills and knowledge about managing storage and virtual networks on Azure, this training is what you are looking for! Train and get the official certification «**Microsoft Certified Azure Administrator Associate**».

DISCOVER THE CURRICULUM

POWER BI

Do you want to provide a complete analysis of a dataset and improve your dashboarding skills? This training is for you! Learn to master Power BI and get your official Microsoft certification by becoming a «**Power BI Data Analyst Associate**».

DISCOVER THE CURRICULUM

DEVELOPPING ON AWS

If you want the ability to write and deploy cloud-based applications, look no further! This course is for you. Upon completion you will be certified as an **AWS Certified Developer Associate**.

DISCOVER THE CURRICULUM

ARCHITECTING ON AWS

If you want to know how to make architectural decisions in accordance with AWS, this is the certification for you! Earn the status of «**AWS Certified Solutions Architect Associate**».

DISCOVER THE CURRICULUM

DataScientest

DO YOU WANT TO BECOME A DATA ENGINEER ?

datascientest.com