SYLLABUS DATA ENGINEER

Master Big Data and the implementation process

DataScientest

SUMMARY

DATASCIENTEST IN NUMBERS

We are DataScientest. Not only are we the market leader in Data Science training in France. We also are the learning institute who offer business-oriented training to professionals and individuals who want to upgrade their skills in the data field.

+70	**Fortune Global 500** Companies trained
+15k	**Alumni**
+3k	Hours of **content**

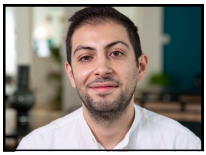
Key figures for the Data Engineer in 2023

261	Number of **subscribers**
80,2%	**Satisfaction** rate

They trust us

OUR TEAM

DataScientest's teaching team is made exclusively of internal professors. They are fully dedicated to teaching and are consistently updating our various training and expert courses.



Charles S.

CTO & Academic Manager
(9 *years of experience)*

Having graduated of the "École Polytechnique", Charles is specialized in programming, Machine Learning and Deep Learning. As CTO of DataScientest, he leads both the faculty and developers working on the platform.

Raphaël K.

Pedagogical Director
(9 years of experience)

Raphaël has a Master's degree in "Statistical Learning and Data Science" from the University of Paris-Dauphine. He designed the Data Analyst course thanks to his knowledge of programming, data visualization and machine learning.

Lecturers

The Data Engineer course is run by lecturers with degrees from Frence´s leading universities. They are selcted for their expertise and teaching skills.

Dan C.

Data Engineer training coordinator- 4 years experience

Dan holds a Master's degree in Mathematical Engineering from the University of Paris Descartes and has extensive experience in consulting. He specialises in Big Data and is now the head of our Data Engineer course.

Dimitri C.

Data Engineer - 3 years experience

Not only is Dimitri a Google Cloud Authorized Trainer, but he holds a Master's degree from the Ecole Normale Supérieure in Mathematics and Computer Vision and a master's from ENSAE in Applied Mathematics. After that he specialised in Data Engineering.

Antoine B.

Data Engineer - 2 years experience

After studying at Ecole 42, a computer programming school, Antoine decided to specialise in data and joined DataScientest as a Data Engineer.

Book an appointment

OUR PEDAGOGICAL APPROACH

DataScientest offers you a course which is a 100% in English and distance learning with a pedagogy based on Learning By Doing.

Training Goals

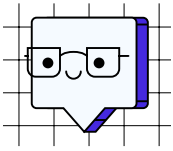
- **Developing** the skills needed to carry out Data Engineering activities
- **Deploying** an artificial intelligence solution within an organisation

Hybrid Format

- **Practical work:** For 80% of your time, you are going to be on our personalized learning platform **Learn**, which was developed by DataScientest. The platform includes online exercises enabling you to gradually implement the concepts developed in the course.
- **Masterclass:** 20% of the training take place in Masterclasses. These face-to-face live-web training sessions with our professors allow you to address all the current problems relating to technologies, methods and tools in the field.
- **Project-based teaching:** In the **backbone project,** you are going to directly apply your newly acquired skills.

Support and assistance

Every day of the week from 9am to 5pm, all the data expert take turns on a dedicated forum to offer personalized technical assistance to you. Educational support is also offered via the "**Slack**", our communication network, and throughout our Q&A sessions.



"We quickly realised that DataScientest had the same vision of teaching and learning as Orange and that they would be a partner that would listen to our specific needs."

Anne Beaugendre-Frénot
Director, Orange Campus Data IA @Orange

Backbone project

Right from the beginning of your training, you are going to carry out a concrete project. The goal of the project is to gradually combine all the skills you obtain during your course. It requires an investment of about 150 hours of work throughout the training.

You will be able to select a project from our list in groups of two or four people. Our topics are updated monthly and are inspired by the work we do in companies. You can also propose a personal project, as long as the data is accessible, and our teaching team confirms it!

This part of our course is crucial to make you fully operational as a Data Engineer. You are going to work on uncleared data sets and produce a professional project. Some coaching sessions are regularly organized by your project mentor in order to guide and coach you.

This allows you to move efficiently from theory to practice and ensures 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 in Data Engineer with a successful project during your job interviews.

If you want to know more about our learners' projects, we have created Data Days, a live broadcast of their projects.

Check the Data Days replay

"DataScientest has genuine expertise in Data Science, delivered with tailor-made support and a constant focus on customer satisfaction."

Xavier Bocher
Head of Credit Risk Internal Models & Operational Research @Groupe Crédit Agricole

Evaluation methods

An evaluation system has been set up throughout your training process. From your first meeting with a DataScientest advisor, we **collect your expectations and needs** at registration. Then, before validating your entry into the course, a **placement test** is sent to you and your knowledge is evaluated using the exams at the end of each module. Finally, through **satisfaction questionnaires**, we collect your appreciation and reviews.

Certification training: Academic recognition

Our certification is issued by The Sorbonne University in Paris. By completing our Data Engineer training, you will receive an official certificate of the French university, Paris-Sorbonne. This will greatly enhance your resume for future job applications.

Book an appointment

Prerequisites

In order to join the Data Engineer course, it is of advantage to have a bachelor's degree. Good knowledge in SQL and Python language, as well as the Linux system, are advantageous.
To take the course, the student must have a computer with an internet connection and a webcam

In Bootcamp or Continuous

Two formats are possible for the Data Engineer training:

Bootcamp

Train quickly, following an intensive 13-week program.

- **Duration :** 13 weeks
- **Pace :** 35-40h/week
- **Total duration :** 500h

Continuing format

Flexible, following the training program while continuing to work.

- **Duration :** 9 months
- **Pace :** 10-12h/week
- **Total duration :** 400h

Book an appointment

THE CURRICULUM

1 - Programming
Python, Advanced Python , Web Scraping

2 - Advanced tools
Bash, Git & Github

3 - Big Data Variety
SQL, MongoDB, ElasticSearch, Neo4j, Hbase

4 - Batch & Streaming
PySpark, Streaming with Spark, Kafka

5 - Practical Data Warehousing
Snowflake, Data Warehousing with DBT (ELT)

6 - Cloud AWS
AWS Solutions Architect

7 - Machine Learning
Statistiques, Data Visualisation, Machine Learning, MLFlow

8 - DevOps - Virtualisation
Docker, API et sécurisation, Kubernetes

9 - CI/CD and Monitoring
Unit Testing with Python, Airflow, GitLab, Prometheus & Grafana

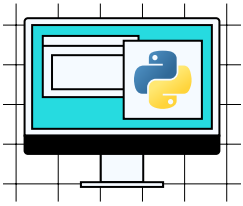
Programming- *duration 40h*

Python

- Mastery of variables and data types
- Overview of various operators and their applications
- Introduction to the concept of loops and control structures
- Defining a function in Python and exploring their applications
- Introduction to classes and modules
- Preparation for the setup, configuration, and chaining of decorators
- Differentiation and implementation of multithreading and multiprocessing in Python
- Applying an asynchronous function in Python
- Introduction to type annotations and use of the MyPy library

Web Scraping

- Introduction to Web Scraping with Beautiful Soup
- Learning how to navigate an HTML document and identify page data



Skills acquired upon completion

- Master the Python language and all its applications
- Understand and use object-oriented programming
- Create complex scripts with Python
- Automatically collect data from a web page

Advanced tools - *duration 20h*

Git

- Introduction to the Git version control system
- Initializing a Git repository
- Presentation and in-depth exploration of Git concepts:
  - Branches
  - Tags
  - Merge

GitHub

- Introduction to the GitHub platform for collaborative work with Git
- Overview of GitHub's key features:
  - Fork
  - Pull Request
  - Issues
- Sharing changes using pull and push
- Introduction to GitHub Actions with practical examples

Système Linux and Script Bash

- Introduction to Linux systems
- Getting started with and using a terminal
- Creating and running Bash scripts

Skills acquired upon completion

- Master version control tools
- Collaborate effectively and version projects using Git and GitHub
- Be able to implement unit tests
- Apply appropriate methods depending on different challenges
- Master the Linux operating system
- Learn how to use a terminal
- Create and manage Bash executables

Big Data Variety - *duration 50h*

SQL

- Introduction to relational databases
- Introduction to the basics of the SQL language
- Deeper understanding of SQL and its applications

Neo4J (optional)

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

MongoDB

- Introduction to MongoDB
- Getting to grips with MongoDB query syntax

Elasticsearch (optional)

- Description of a search engine
- Presentation of an index and how to use it
- Mapping development
- Pre-processing data with Ingest Node
- Data extraction with Text Analyzer

HBase (optional)

- Introduction to column-oriented databases
- Querying and modifying data with Python and Happybase

Skills acquired at the end

- Knowing how to choose a database management system depending on the use case
- Understanding how to query an RDBMS (relational database management system) using the SQL language
- Handling a document-oriented database such as MongoDB
- Improving the search of your textual data using Elasticsearch
- Managing a graph-oriented database
- Using Cypher to query and update graph-oriented databases
- Initiation and queries with a column-oriented database

Batch & Streaming - *duration 50h*

PySpark

- Introduction to distributed computing with PySpark
- Introduction to Spark's RDD and Dataframe APIs
- Distributed data processing pipeline with PySpark
- Distributed Machine Learning with Spark MLLib

Kafka

- Presentation of the Kafka distributed streaming platform:
  - Architecture
  - Advantages
- Management of Producer settings (partition key)
- Control of Consumers settings (Consumer group)

Streaming with Spark

- Getting to grips with Spark Streaming for real-time data processing
- Presentation of the mini-batch streaming required to run Spark Streaming
- Overview of Structured Streaming features
- Creating a processing pipeline with Structured Streaming

Skills acquired at the end

- Grasp the fundamental concepts of Big Data
- Understand the theory behind distributed system architectures
- Learn how to manage real-time data streams
- Process and transform data in real time using distributed computing with Spark Streaming

Practical Data Warehousing – *Duration: 20 hours*

Snowflake

- Data Warehousing with robust security and compatibility with major cloud providers
- SQL-based data storage and analysis built for the cloud
- Offers an architecture that separates storage and computation to optimize costs and performance.

Data Warehousing With DBT (ELT)

- Transformations from Data Warehouses
- Producing high-quality datasets
- Automating the execution of transformations and generating clear, interactive documentation.

Skills acquired at the end

- Use Snowflake for SQL data storage and analysis in a cloud environment.
- Perform data transformations from data warehouses and produce high-quality datasets.
- Create and deploy ETL workflows and generate data profiles to ensure data transparency.

Cloud AWS - *duration 70h*

AWS Solutions Architect

- Fundamentals and Best Architectural Practices on the Cloud
- Designing highly available and resilient architectures on AWS
- Continuous improvement and automation of architecture deployment
- Overview of AWS cloud and the global infrastructure basics
- Key services of the AWS platform and their common use cases

Skills acquired at the end

- Understand the features and use cases of AWS Cloud services (EC2, EBS, ELB, EventBridge, ECS/EKS, RDS & DynamoDB)
- Leverage AWS cloud services
- Deploy and monitor infrastructure and applications on AWS cloud

Machine Learning - *duration 60h*

Machine Learning

- Data pre-processing
- Selection and optimisation of a Machine Learning algorithm
- Definition and application of a regression algorithm
- Definition and application of a classification algorithm
- Development of clustering algorithms

MLflow

- Introduction to the MLFlow Architecture
- MLFlow Tracking, MLFlow Projects, MLFlow Models, MLFlow Registry
- Managing the lifecycle of a Machine Learning project

Data Visualisation

- Introduction to different types of graph with MatplotLib:
  - Barplots
  - Scatter plots
  - Histograms
  - Box plots
  - Pie plots
- Dash applications

Statistics

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

Skills acquired at the end

- Understand the fundamentals of key Machine Learning algorithms
- Be immediately operational in machine learning
- Train machine learning models using the Scikit-Learn library
- Manipulate your data with Pandas DataFrames
- Master Numpy
- Visualize your data with various charts using Matplotlib
- Master the MLFlow architecture

DevOps - Isolation - *duration 50h*

APIs

- Introduction to APIs and Microservices Architectures
- Overview of different HTTP methods and their functions
- Using FastAPI and Flask libraries to develop RESTful APIs
- Documenting an API with the OpenAPI specification
- Managing errors and performance of an API

Securing APIs

- Introduction to API security and API keys
- Basic HTTP authentication
- Introduction to JSON Web Token and HTTPS

Docker

- Introduction to containerisation and its usefulness in relation to virtualisation
- Introduction to how Docker works
- Handling images and containers
- Communicating with containers
- Data persistence using volumes
- Creating a Docker image using a Dockerfile
- Sharing images on Dockerhub
- Using docker-compose

Kubernetes

- Deploy and Manage Containers
- Initialization and architecture
- Deploying an API with Kubernetes

Skills acquired at the end

- Understand APIs
- Learn how to create an API with Flask and FastAPI
- Make HTTP API requests
- Understand virtualization
- Master containerization techniques and tools
- Master container orchestration with Kubernetes
- Deploy pods using the Kubernetes interface
- Secure APIs

CI/CD and Monitoring- *duration 40h*

Airflow

- Introduction to Airflow Concepts
- Overview of orchestration principles and their usefulness
- Directed Acyclic Graphs (DAGs)
- Operators
- Task management through specific Operators
- Monitoring DAGs via the Airflow graphical interface

Gitlab

- Installation, Initialization, Adding and Removing Documents
- Git Blame, Tag
- Checking the status of your local repository
- Branching, creating, switching, merging
- Conflict management

Tests unitaires avec Python

- Setting up unit tests with Pytest
- Introduction to integration tests and their functions
- Overview of the benefits of testing
- Integrating unit tests into a development environment

Prometheus & Grafana

- Understand the importance of monitoring
- Using Prometheus Query Language
- Creating Dashboards with Grafana
- Integration into a production environment

Les compétences acquises à l'issue

- Automate your processes by mastering Apache Airflow
- Trigger alternative tasks in case of failures
- Verify the functionality of independent code units during development
- Integrate a monitoring tool into a production environment

OUR PARTNERS

DataScientest has developed partnerships with world-renowned institutions. On one hand, with academic institutions such as Université Paris - Panthéon Sorbonne, and on the other hand, with software publishers such as AWS or Microsoft. These partnerships are designed to help learners distinguish themselves from other candidates by obtaining certifications recognized by companies.

Amazon Web Services

Software partner

Today, DataScientest enjoys the exclusive status of an Amazon Digital Partner. Therefore, we have been authorized by Amazon to train teams on their products and services. As part of this partnership, we have built several courses that prepare you for the official **AWS certifications,** such as **Cloud Practitioner**. The registration fee for the official exam is included in the course price.

Microsoft

Software partner

DataScientest is a **Microsoft Learning Partner** and therefore authorized to train you in official Microsoft certifications (**PL-900 or AZ-900**). These certifications attest to a certain level of expertise in Azure, the collection of cloud computing products and services, and in Power BI, Microsoft's business intelligence tool.

University Paris 1 - Panthéon Sorbonne

Academic partner

Our **certification** is issued by the **Sorbonne University** in Paris. By completing our Data Analyst, Data Scientist, Data Engineer or DevOps training, you will receive an official certificate of the French university Paris-Sorbonne. This will greatly enhance your resume for future job applications.

ALUMNI TESTIMONIES

Steve NICOLE

Structural Engineer @Framatome

"I did my training to become a Data Engineer at DataScientest. The training material is of a high standard and the teaching team is committed, responsive and concerned about the success of its learners. Data engineer training is very demanding but rewarding - I recommend it!!!!"

Sajida Shaik

Diligent Software Engineer @Stellantis

"The training is designed, progressing from basic to advanced level, and it is more practical than theoretical, This practical approach has significantly enhanced my problem-solving skills for real-world projects. The availability of a forum for posting doubts has been particularly helpful in ensuring a smooth learning experience."

Sofian GAIDE

Data Engineer Intern @Ippon Technologies

"3 months of intense but extremely rewarding training! See Data from A to Z so that you're ready to work with all kinds of data in your company! I highly recommend this course!"

Hana MOUJOU

"The course is rich of information, exciting and very demanding! There's a lot of pedagogy in the online courses and masterclasses, as well as a real coherence in the educational path. This is followed by Data Engineering projects on various DevOps and Big Data technologies. The teaching staff are friendly, available and responsive. In short, a course that I can highly recommend!"

Customer Care

The Customer Care team is made up of **technical advisors** and program managers who work together to provide supervised **support** to each class. The technical side of the training is managed by the technical advisors, while the human side is taken care of by the programme managers.

The classes and learners are **monitored individually** to help them achieve their diploma. To this end, we organize training **follow-up interviews** which emphasize the human dimension and allow us to adapt the rules in the event of personal complications. We also offer **advices** on how to manage time and improve the way you learn throughout the course. **Events** are also organized to strengthen group cohesion and avoid the isolation of distance learners.

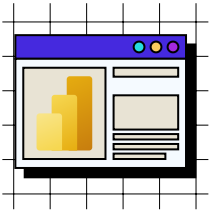
We are attentive to the comments of our users to continue to improve our training programmes through satisfaction questionnaires. In addition, there are profiles dedicated to supporting and dealing with each learner's questions.

"Caring for our learners is our main goal. We make sure that their training goes well through individual follow-up, advice and an attentive ear to the needs of each individual."

Pauline Messager
Head of Customer Care Service @DataScientest

TO GO FURTHER

After your training, if you want to build on your skills, DataScientest has set up a range of publisher certifications, such as Microsoft and AWS, so you can deepen your knowledge and perfect your Data skills!



Power BI

You want to provide a complete analysis of a dataset and improve your dashboard creation? This course is for you! Learn to master Power BI and earn your official Microsoft certification by becoming a "**Power BI Data Analyst Associate**".

Discover the curriculum »

DataScientest

DO YOU WANT TO BECOME A DATA ENGINEER?

datascientest.com

contact@datascientest.com

+49 32 222003762