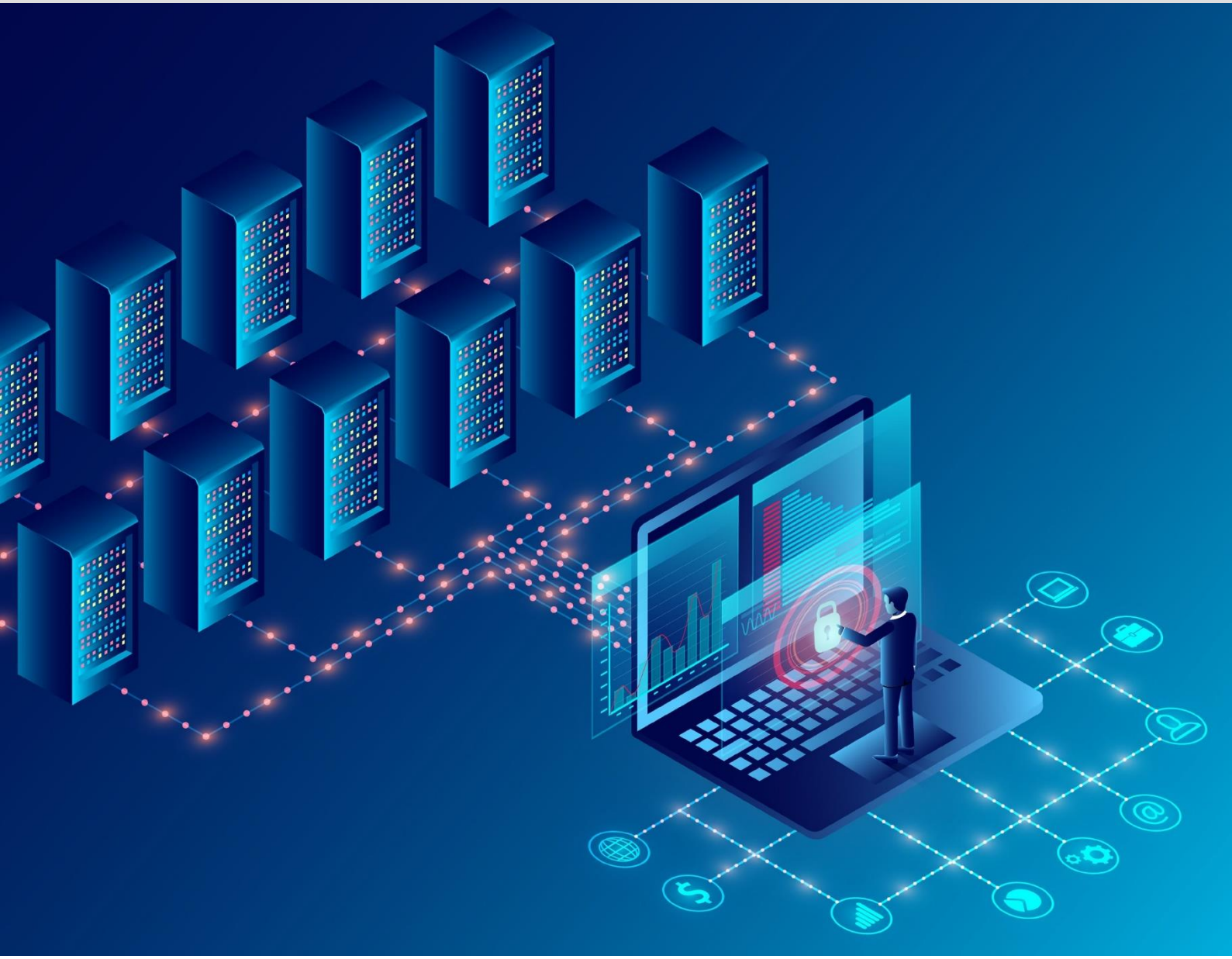
DE LORENZO

DL NGL

NEXT GENERATION LABS



DL NGL-DATA

Module for the study of big data

DE LORENZO

DL NGL

The 'Big Data' module deals with the study of the processing and analysis of large amounts of data in the context of Data Science.

Big Data differs from traditional data collections in several characteristics: the amount of data, the fact that data is generally unstructured because it comes from different sources and forms and, in the case of real-time streaming, the speed with which the data arrives.

New technologies have been introduced in Data Science that deal with the management and analysis of large data, overcoming the limitations of traditional data management systems such as relational DBMS (Database Management System).

The ‘Big Data’ module uses Apache Spark, an open-source framework that supports in-memory parallel computing to optimize the performance of applications that analyse Big Data.
It is used by a great many organizations around the world, including IBM, NASA, Samsung and Yahoo!, and its use is constantly expanding.

Within the Module, simple applications will be developed using the Python programming language.

The 'Big Data' module addresses the main issues of big data processing with an informative approach and with practical exercises that make learning more effective.
The data that is analysed can be of different types: structured, semi-structured or unstructured. The most common structures in Apache Spark are: DataFrame, dataset and RDD (Resilient Distributed Dataset).

The analysis of Big Data often has the objective of making predictions or deriving any strategies for users based on the available data.

For advanced analysis and Machine Learning-type processing, the Spark MLib library is used, which allows performing preliminary data analyses, implementing regression and classification models, making predictions, but also clustering.
In fact, it provides various algorithms, such as those for linear regression, the K-means method, the random forests, etc.

The Module will examine different processing techniques with different purposes, for example, Supervised Learning in which, starting from historical data, we will try to train a model to predict future data through an iterative optimization algorithm.

DE LORENZO

DL NGL

Or, the Unsupervised Learning, where you try to find patterns or discover the underlying structure in a set of data.

This methodology is particularly indicated in the search for anomalies, but it is also very useful in marketing, to carry out segmentation and, therefore, identify groups that are as homogeneous internally and heterogeneous as possible.

Finally, the processing of data represented by graphs, graph operators, graph algorithms, graph builders and other techniques are examined. A graph is a set of nodes (or vertices) connected to each other through arcs and can be used to represent, for example, the friendship network in a social network where the nodes are people while the arcs represent friendships.

Spark's GraphX library provides algorithms for measuring the importance of a node, the presence of any hubs, calculating connected components, calculating triangles, etc.
Most of the practical exercises will be carried out on real data taken from the site www.kaggle.com, which contains datasets accessible to all and of different formats.

Educational experience

- Introduction to Big Data
- Concepts of database and statistics
- Architecture of a Big Data analysis platform
- Parallel data processing
- Use of Data Frames
- SQL queries on structured data
- Models of advanced analysis and Machine Learning
- Processing of data represented by graphs
- Examples and applications

NEXT GENERATION LABS

The DL NGL-DATA module can be integrated in the NEXT GENERATION LAB - DL NGL laboratory through the minimum purchase of the following modules:

- **Teacher Station - DL NGL-BASE**
  Necessary for the proper functioning of the laboratory. Quantity: 1.

- **Student Station - DL NGL-STUDENT**
  To be multiplied by the number of "student stations" to be created.