Sources of data are heterogeneous, flow at different speeds and vary in volume. Defining a strong data integration framework is key to report on data quality and to efficiently explore and visualize their contents. Big Data can be characterized by the 3 Vs of Volume, Velocity and Variety and often requires dedicated computing solutions, which will be explored.

Prerequisites

- Computer science and programming skills (mainly Python language: numpy and pandas libraries).
- Work experience of minimum 3 years.

Learning objectives

After completing this course, participants will be able to:

- Explain the key components of ETL-based data warehousing;
- Set up indicators on data quality and management;
- Perform a simple data visualization task;
- Implement the distribution of simple operations via the Map/Reduce principle in Spark;
- Connect on a cloud computing engine (e.g. Google Cloud Platform) and launch a simple task;
- Deploy a Docker container.

Information and registration

info.exed@isae-supaero.fr
Course Content

AIBT1 – Data integration and processing for value creation
From the Advanced Master AIBT
(Artificial Intelligence and Business Transformation)

AIBT101 – Introduction to modern AI (28h):
- AI Basics;
- Landscape and flagship algorithms on Supervised;
- Unsupervised and Reinforcement Learning;
- Understanding the relationship between problem framing;
- Types of data available;
- Actual business outcomes and the applicable algorithms;
- Business intelligence and business models;
- Major success stories of Business and AI;
- Google’s Self-driving car; IBM Watson’s Medical diagnosis;
- DeepMind’s Alpha Go beating the World champion of Go;
- Airbus building the Skywise platform;
- AI to deliver prescription for manufacturing;

AIBT102 – Data integration and exploration (28h):
Data Warehousing:
- History and recent evolutions
- Extract-Transform-Load process
- Architecture
- Key functions
- Layers

Data quality:
- Indicators
- improvement
- assurance
- control

Data visualization:
- visual perception
- effective graphical display
- tools

AIBT103 – Big data processing (28h):
Distributed computing with Spark:
- History
- MapReduce paradigm
- Hadoop Stack
- Hadoop Distributed File System
- MLlib Machine Learning library
Virtualization and cloud computing:
- Different approaches to virtualization
- Economical models
- Technical benefits (snapshots, dynamic deployment and migration, failover...)
- Cloud engines (principles, deployment examples, node choices)

Docker:
- Fundamental differences w.r.t. virtualization
- Docker components
- Tools

Teaching methods

<table>
<thead>
<tr>
<th>Teaching methods</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectures / tutorial</td>
<td>X</td>
</tr>
<tr>
<td>Collaborative learning</td>
<td></td>
</tr>
<tr>
<td>Flipped classroom</td>
<td></td>
</tr>
<tr>
<td>Blended learning (online and face to face)</td>
<td></td>
</tr>
<tr>
<td>Learning by doing</td>
<td>X</td>
</tr>
<tr>
<td>Project-based</td>
<td></td>
</tr>
<tr>
<td>Simulation</td>
<td></td>
</tr>
<tr>
<td>Case study</td>
<td>X</td>
</tr>
</tbody>
</table>

Assessment

Written examination (100%)