# Big Data: State of the art

*Seminar of 2 days - 14h*
*Ref.: BGA - Price 2025: 2 140 (excl. taxes)*

## EDUCATIONAL OBJECTIVES

At the end of the training, the trainee will be able to:

Learn the main concepts of Big Data.

Identify the economic issues

Evaluate the pros and cons of Big Data.

Understand the main problems and potential solutions

Identify the main methods and areas of application for Big Data

## THE PROGRAMME

last updated: 02/2024

### 1) Introduction

- The origins of Big Data: A world of digital data, e-health, timeline.
- The four-V's definition: Origins of the data.
- A breakthrough: Changes in quantity, quality, and habits.
- The value of data: A change in importance.
- Data as a raw material.
- The fourth paradigm of scientific discovery.

### 2) Big Data: Processing, from acquisition to result.

- The sequence of operations. Acquisition.
- Data collection: crawling, scraping.
- Managing event flows (Complex Event Processing, CEP).
- Indexing incoming flows.
- Integration with old data.
- Data quality: A fifth V?
- Different types of processing: Searching, learning (Machine Learning, transactional learning, data mining).
- Other sequencing models: Amazon, e-Health.
- One or more data repositories? From Hadoop to the in-memory.
- From tonal analysis to knowledge discovery.

### 3) Relationships between the Cloud and Big Data

- The architecture model of public and private Clouds.
- XaaS services.
- The goals and benefits of Cloud architectures.
- Infrastructure.
- Similarities and differences between the Cloud and Big Data.
- Storage clouds.
- Classification, security, and privacy of data.
- Structure as a classification criterion: Unstructured, structured, semi-structured.
- Classification by life cycle: Temporary or permanent data, active archives.
- Security difficulties: Increased volumes, distribution.
- Potential solutions.

### TRAINER QUALIFICATIONS

The experts leading the training are specialists in the covered subjects. They have been approved by our instructional teams for both their professional knowledge and their teaching ability, for each course they teach. They have at least five to ten years of experience in their field and hold (or have held) decision-making positions in companies.

### ASSESSMENT TERMS

The trainer evaluates each participant's academic progress throughout the training using multiple choice, scenarios, hands-on work and more.
Participants also complete a placement test before and after the course to measure the skills they've developed.

### TEACHING AIDS AND TECHNICAL RESOURCES

• The main teaching aids and instructional methods used in the training are audiovisual aids, documentation and course material, hands-on application exercises and corrected exercises for practical training courses, case studies and coverage of real cases for training seminars.
• At the end of each course or seminar, ORSYS provides participants with a course evaluation questionnaire that is analysed by our instructional teams.
• A check-in sheet for each half-day of attendance is provided at the end of the training, along with a course completion certificate if the trainee attended the entire session.

### TERMS AND DEADLINES

Registration must be completed 24 hours before the start of the training.

### ACCESSIBILITY FOR PEOPLE WITH DISABILITIES

Do you need special accessibility accommodations? Contact Mrs. Fosse, Disability Manager, at psh-accueil@ORSYS.fr to review your request and its feasibility.

## 4) Introduction to Open Data

- Philosophy of open data and goals.
- Releasing public data.
- Implementation difficulties.
- Essential features of open data.
- Areas involved. Expected benefits.

## 5) Equipment for storage architectures

- Servers, disks, networks, and use of SSD drives, importance of network infrastructure.
- Cloud architectures and more traditional architectures.
- Benefits and difficulties.
- The TCO. Power consumption: Servers (IPNM), drives (MAID).
- Object storage: principle and benefits.
- Object storage compared to traditional NAS and SAN storage.
- Software architecture.
- Storage management location levels.
- Software-Defined Storage.
- Centralized architecture (Hadoop File System).
- Peer-to-peer and hybrid architectures.
- Interfaces and connectors: S3, CDMI, FUSE, etc.
- Future of other storage types (NAS, SAN) relative to object storage.

## 6) Data protection

- Preservation over time in the face of increased volumes.
- Online or local backups?
- Traditional archiving and active archiving.
- Links with storage hierarchy management: Future of magnetic tape.
- Multisite replication.
- Damage to storage media.

## 7) Scope processing methods

- Classification of analysis methods based on data volume and processing power.
- Hadoop: The Map Reduce processing model.
- The Hadoop ecosystem: Hive, Pig. The difficulties of Hadoop.
- OpenStack and the Ceph data manager.
- Complex Event Processing: An example? Storm.
- From BI to Big Data.
- Return to decisional and transactional models: NoSQL databases. Types and examples.
- Data ingestion and indexing. Two examples: Splunk and Logstash.
- Open-Source crawlers.
- Search and analysis: Elasticsearch.
- Learning: Mahout. In-memory.
- Visualization: Real-time or not, in the Cloud (Bime), comparison of QlikView, Tibco Spotfire, and Tableau.
- A general architecture of data mining via Big Data.

## 8) Usage case through examples and conclusion

- Anticipation: Needs of users within companies, equipment maintenance.
- Security: People, fraud detection (mail, taxes), the network.
- Recommendation. Marketing analysis and impact analyses.
- Path analyses. Distribution of video content.
- Big Data for the automotive industry? For the oil industry?
- Should you begin a Big Data project?
- What future is there for data?
- Governance of data storage: Roles and recommendations, Data Scientists, skills involved in a Big Data project.

## DATES

REMOTE CLASS
2025 : 30 sept., 16 déc.