Course : Real-time Big Data processing with Spark and Storm

Practical course - 3d - 21h00 - Ref. DSS
Price : 2010 € E.T.

Big Data, known for its ability to process massive data, is now integrating a real-time component with the arrival of tools such as Spark and Storm. You'll discover the advantages of these tools, their real-time distributed computing systems and the notion of Streaming Big Data.

Teaching objectives

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

- Understand the fundamentals of real-time big data application development
- Evaluating Spark and Storm
- Applying Storm and Spark real-time distributed computing systems
- Process large quantities of data in real time

Intended audience
Designers, developers, architects.

Prerequisites
Good knowledge of software development. An understanding of big data issues is a plus.

Course schedule

1 **Introduction to real-time architecture**

- Real-time processing.
- Lambda architectures.
- Kappa architectures.
- SMACK architectures.

Hands-on work
Study of the implementation of a Kappa architecture for Spark and Strom.

2 Kafka architecture

- The Kafka Producers, Brokers, Consumers overview.
- Kafka's log files.
- Avro schematics. Using ZooKeeper.

Hands-on work
Study of Kafka configuration in Kappa architecture.

3 Apache Storm architecture

- Definition of the development environment.
- Creation of Storm-based projects.
- Definition of Storm components (Spout and Bolt).
- Storm flow definition.
- Data model (key, value).
- Nimbus and ZooKeeper roles.

Case study
Kappa architecture implementation study for Storm.

4 Handling Storm messages

- Service programming with Clojure, Java, Python.
- Message life cycle.
- The Storm API for defining reliability.
- Reliability implementation strategy for a Big Data application.

Hands-on work
Implementation of a real-time social network processing project in the Kappa architecture.

5 Apache Spark architecture

- The different versions of Spark (Scala, Python, R and Java).
- Comparison with the Storm environment.
- The different Spark modules.
- The different types of architecture: Standalone, Apache Mesos or Hadoop YARN.

Hands-on work
Study of the implementation of the SMACK architecture for Spark.

6 Real time with Spark Streaming

- Introducing Resilient Distributed Dataset (RDD) ?
- Create, manipulate and reuse RDDs.
- Accumulators and broadcast variables.
- Operating principle.
- The different types of source.
- Comparison with Apache Storm.

Hands-on work
Implementation of a real-time social network processing project.

7) Other market players

- Comparison of all streaming tools in the ecosystem (Storm, Spark Streaming, Flink, Samza).
- Focus on Samza architecture.

Hands-on work
Study of the implementation of Kappa architecture with Samza.