Course : Talend Open Studio, data integration for big data

Community version (free and open source)

Practical course - 3d - 21h00 - Ref. IDB

Price : 2010 € E.T.

Talend's data integration platform extends its capabilities to big data technologies such as Hadoop (HDFS, HBase, HCatalog, Hive and Pig) and the NoSQL databases Cassandra and MongoDB. This course will provide you with the basics for using Talend components designed to communicate with big data systems. This course is exclusively about Talend Open Studio (free, open source, community version). It does not cover the commercial version of Talend Studio, under paid license, integrated into the Qlik-Talend Cloud portal.

Teaching objectives

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

- Mastering Talend in a big data environment
- Use Talend as a link between files, applications and databases
- Acquire the tool's philosophy
- Adopt best practices and design flexible, robust information systems
- Be able to implement jobs
- Read and write data to HDFS and NoSQL databases with Talend jobs
- Transforming jobs with Pig and Hive
- Managing data quality with Talend
- Using Sqoop to facilitate the migration of relational databases to Hadoop
- Master the use of the component library
- Perform simple and complex end-to-end ETL (Extract, Transform and Load) processing

Intended audience

BI consultants, architects, project managers, data managers or anyone who needs to manage data flows.

Prerequisites

Knowledge of Hadoop, Spark and Kafka.

PARTICIPANTS

BI consultants, architects, project managers, data managers or anyone who needs to manage data flows.

PREREQUISITES

Knowledge of Hadoop, Spark and Kafka.

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.

Practical details

Succession of mini-projects leading to the design of Talend big data jobs of increasing difficulty.

Course schedule

1 Talend Open Studio presentation

- Data integration. ETL solutions.
- Big data. Unstructured data. NoSQL databases.
- The Hadoop ecosystem (HDFS, MapReduce, HBase, Hive, Pig, etc.).
- TOS for Data Integration: data integration.
- TOS for Data Quality: data quality management.
- TOS for big data.
- Product philosophy.

Hands-on work

Installation/configuration of TOS for big data. Getting started.

2 Designing jobs

- Introduction to business modelers and job designers.
- Simple transformation components.
- View generated code, run a job.
- Set up jobs.
- Create and manage your own variables.
- Good design practices.

Hands-on work

Development of a job connecting to a data source, filtering, transformation and storage of the result in a file.

3 Data integration in cluster and NoSQL databases

- Definition of Hadoop cluster connection metadata.
- Connect to a MongoDB, Neo4j, Cassandra or Hbase database and export data.
- Simple data integration with a Hadoop cluster.
- Presentation of extension components.
- Utilisation du composant d'extension : capture de tweets et importation directe dans HDFS.

Hands-on work

Read tweets and store them as files in HDFS, analyze the frequency of topics and store the results in HBase.

4 Import/export with Sqoop

- Use Sqoop to import, export and update data between RDBMS and HDFS systems.
- Partial, incremental import/export of tables.
- Import/export a SQL database to and from HDFS.
- Big data storage formats (AVRO, Parquet, ORC, etc.).

Hands-on work

Migrate relational tables to HDFS and vice versa.

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.

5 Manipulate data

- Introducing the Pig brick and its PigLatin language.
- Talend's main Pig components, Pig flow design.
- Development of UDF routines.

Hands-on work
Identify website usage trends by analyzing logs.

6 Architecture and best practices in a Hadoop cluster

- Design efficient storage in Hadoop.
- Data lake versus data warehouse: should you choose?
- Hadoop and the Disaster Recovery Plan (DRP) in the event of a major incident.
- Automate workflows.

Hands-on work
Create your own data lake and automate its operation.

7 Analyze and store your data with Hive

- Hive connection and schema metadata.
- The HiveQL language.
- Hive flow design, query execution.
- Implement Hive's ELT components.

Hands-on work
Store stock price trends in HBase and consolidate them with Hive, so as to materialize hour-by-hour trends for a given day.

Dates and locations

REMOTE CLASS
2026 : 29 June, 18 Nov.

PARIS LA DÉFENSE
2026 : 22 June, 4 Nov.