Course : Hadoop Cloudera developer, certification preparation (CCA175)

Practical course - 4d - 28h00 - Ref. HDZ

This course will teach you the key concepts and expertise needed to integrate and store data in a Hadoop cluster using the latest techniques and tools. It prepares you for the certification "CCA Spark and Hadoop developer".

Teaching objectives

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

- Discover the Hadoop ecosystem
- Understand the HDFS distributed file system and master MapReduce processing and code writing
- Best practices for developing and implementing common algorithms
- Optimize configurations and improve performance
- Using Hive, Pig, Flume, Mahout and Sqoop for Hadoop ecosystem projects
- Preparing for Cloudera certification

Intended audience

Project managers, developers, data scientists, and anyone wishing to understand development techniques using MapReduce in the Hadoop environment.

Prerequisites

Basic knowledge of an object-oriented programming language.

Certification

Following the course, you can take the Cloudera Certified Associate Spark and Hadoop Developer (CCA175) exam. This exam takes place outside the training course. The aim is to become a certified Cloudera expert in your company. To register, visit www.examslocal.com.

Practical details

Teaching methods
This big data training course includes 50% practical work over the 4 days.

Course schedule

① Hadoop, HDFS and distributed processing on a Hadoop cluster

- General introduction to Hadoop and its ecosystem.
- Data processing.
- HDFS: the Hadoop file system.
- The components of a Hadoop cluster.
- HDFS architecture. Using HDFS.
- YARN architecture and working with YARN.

② Spark basics

- Introduction to Spark.
- Start up and use the Spark console.
- Introduction to Spark Datasets and DataFrames.
- Operations on DataFrames.

③ DataFrame and schema manipulation, data analysis with queries

- Create DataFrames from various data sources.
- Saving DataFrames. DataFrame schemas.
- Gluttonous and lazy execution of Spark.
- Query DataFrames with expressions on named columns.
- Grouping and aggregation queries.
- Joints.

④ RDDs and table and view querying with Spark SQL

- Spark's fundamental structure.
- Transform data with Resilient Distributed Dataset (RDD).
- Data aggregation with pair RDDs.
- Query tables in Spark using SQL.
- Query files and views.
- The Spark Catalog API.

⑤ Working with Spark

- Working with Spark datasets in Scala. The differences between datasets and DataFrames.
- Create, load and save datasets. Operations on datasets.
- Write, configure and run Spark applications.
- Writing a Spark application. Compiling and launching an application. Deploying an application.
- Spark applications web user interface. Configure application properties.
- Distributed processing with Spark. A reminder of how Spark works with YARN.
- Data partitioning in RDDs, queries, jobs, steps and tasks.

⑥ Persistence of distributed data

- DataFrame and dataset persistence.
- Persistence levels.
- Persistent HHW

---

7   Iterative algorithms with Spark and introduction to Spark streaming

- Other common uses of Spark.
- Iterative algorithms in Spark. Machine learning with Spark.
- Introduction to Spark streaming. Creating streaming DataFrames.
- Transform DataFrames. Execute streaming requests.

8   Structured streaming with Kafka and operations on streaming

   DataFrames

- Introduction. Receiving and sending Kafka messages.
- Aggregate and join streaming DataFrames.