

Opleiding : Hadoop Cloudera ontwikkelaar, certificeringsvoorbereiding (CCA175)

Praktijkcursus - 4d - 28u00 - Ref. HDZ

In deze cursus leer je de belangrijkste concepten en expertise die nodig zijn om gegevens te integreren en op te slaan in een Hadoop-cluster met behulp van de nieuwste technieken en tools. De cursus bereidt je voor op de certificering "CCA Spark and Hadoop developer".

Pedagogische doelstellingen

Aan het einde van de training is de deelnemer in staat om:

- ✓ Ontdek het Hadoop-ecosysteem
- ✓ Het gedistribueerde bestandssysteem HDFS begrijpen en MapReduce-verwerking en -code schrijven onder de knie krijgen.
- ✓ Goede praktijken bij het ontwikkelen en implementeren van gemeenschappelijke algoritmen
- ✓ Configuraties optimaliseren en prestaties verbeteren
- ✓ Hive, Pig, Flume, Mahout en Sqoop gebruiken voor Hadoop ecosysteemprojecten
- ✓ Voorbereiden op Cloudera-certificering

Doelgroep

Projectmanagers, ontwikkelaars, datawetenschappers en iedereen die ontwikkeltechnieken met behulp van MapReduce in de Hadoop-omgeving wil begrijpen.

Voorafgaande vereisten

Basiskennis van een object programmeertaal.

Certificatie

Na afloop van de cursus kun je het Cloudera Certified Associate Spark and Hadoop Developer (CCA175) examen afleggen. Dit examen vindt plaats buiten de cursus om. Het doel is om een gecertificeerde Cloudera expert in uw bedrijf te worden. Ga naar www.examslocal.com om je in te schrijven.

DEELNEMERS

Projectmanagers, ontwikkelaars, datawetenschappers en iedereen die ontwikkeltechnieken met behulp van MapReduce in de Hadoop-omgeving wil begrijpen.

VOORAFGAANDE VEREISTEN

Basiskennis van een object programmeertaal.

VAARDIGHEDEN VAN DE CURSUSLEIDER

De deskundigen die de cursus leiden zijn specialisten op het betreffende vakgebied. Zij werden geselecteerd door onze pedagogische teams zowel om hun vakkennis als hun pedagogische vaardigheden voor elke cursus die zij geven. Zij hebben minstens vijf tot tien jaar ervaring in hun vakgebied en oefenen of oefenden verantwoordelijke bedrijfsfuncties uit.

BEOORDELINGSMODALITEITEN

De cursusleider beoordeelt de pedagogische vooruitgang van de deelnemer gedurende de gehele cursus aan de hand van meerkeuzevragen, praktijksituaties, praktische opdrachten, ... De deelnemer legt ook van tevoren en naderhand een test af ter bevestiging van de verworven kennis.

Praktische modaliteiten

Leer methodes

Deze big data training omvat 50% praktisch werk gedurende de 4 dagen.

Opleidingsprogramma

1 Hadoop, HDFS en gedistribueerde verwerking op een Hadoop-cluster

- Algemene inleiding tot Hadoop en het bijbehorende ecosysteem.
- Gegevensverwerking.
- HDFS: het Hadoop-bestandssysteem.
- De onderdelen van een Hadoop-cluster.
- HDFS architectuur. HDFS gebruiken.
- De architectuur van YARN en werken met YARN.

2 Basisprincipes van vonken

- Inleiding tot Spark.
- Start en gebruik de Spark-console.
- Inleiding tot Spark Datasets en DataFrames.
- Bewerkingen op DataFrames.

3 Omgaan met DataFrames en schema's, gegevensanalyse met query's

- DataFrames maken van verschillende gegevensbronnen.
- DataFrames opslaan. DataFrame schema's.
- Gulzige en luie uitvoering van Spark.
- Query DataFrames met expressies op benoemde kolommen.
- Query's voor groeperen en aggregeren.
- Gewrichten.

4 RDD's en query's op tabellen en views met Spark SQL

- Fundamentele structuur van Spark.
- Gegevens transformeren met veerkrachtige gedistribueerde datasets (RDD).
- Gegevensaggregatie met RDD's in paren.
- Tabellen in Spark doorzoeken met SQL.
- Bestanden en weergaven bevragen.
- De Spark catalogus API.

5 Werken met Spark

- Werken met Spark datasets in Scala. De verschillen tussen datasets en DataFrames.
- Datasets maken, laden en opslaan. Bewerkingen op datasets.
- Spark-toepassingen schrijven, configureren en uitvoeren.
- Een Spark-applicatie schrijven. Een applicatie compileren en starten. Een applicatie implementeren.
- De web user interface voor Spark applicaties. De eigenschappen van een applicatie configureren.
- Gedistribueerde verwerking met Spark. Een herinnering aan hoe Spark werkt met YARN.
- Datapartitionering in RDD's, query's, jobs, stappen en taken.

PEDAGOGISCHE EN TECHNISCHE MIDDELEN

- De gebruikte pedagogische middelen en cursusmethoden zijn voornamelijk: audiovisuele hulpmiddelen, documentatie en cursusmateriaal, praktische oefeningen en correcties van de oefeningen voor praktijkstages, casestudies of reële voorbeelden voor de seminars.
- Na afloop van de stages of seminars verstrekt ORSYS de deelnemers een evaluatievragenlijst over de cursus die vervolgens door onze pedagogische teams wordt geanalyseerd.
- Na afloop van de cursus wordt een presentielijst per halve dag verstrekt, evenals een verklaring van de afronding van de cursus indien de stagiair alle sessies heeft bijgewoond.

TOEGANGSMODALITEITEN EN TERMIJNEN

De inschrijving dient 24 uur voor aanvang van de cursus plaatsgevonden te hebben.

TOEGANKELIJKHEID VOOR MINDERVALIDEN

Is voor u speciale toegankelijkheid vereist? Neem contact op met mevr. FOSSE, contactpersoon voor mindervaliden, via het adres psh-accueil@ORSYS.fr om uw verzoek en de haalbaarheid daarvan zo goed mogelijk te bestuderen.

6 Persistentie van gedistribueerde gegevens

- De persistentie van DataFrames en datasets.
- Persistentie niveaus.
- Persistent HHW

7 Iteratieve algoritmen met Spark en inleiding tot Spark streaming

- Andere veelgebruikte toepassingen van Spark.
- Iteratieve algoritmen in Spark. Machinaal leren met Spark.
- Inleiding tot Spark streaming. Aanmaken van streaming DataFrames.
- DataFrames transformeren. Streaming verzoeken uitvoeren.

8 Gestructureerd streamen met Kafka en bewerkingen op streaming

DataFrames

- Inleiding. Ontvangen en verzenden van Kafka berichten.
- Aggregatie en join op streaming DataFrames.