Científico/a de datos Big Data Cloudera (virtual)

Descrición

Os/As data scientists son os/as encargados/as de construír plataformas de información para proporcionar unha visión profunda e responder a preguntas previamente inimaxinables. Spark e Hadoop están a transformar a forma de traballar dos/as data scientists ao permitir a análise de datos interactivos e iterativos a escala.

Aprenda como Spark e Hadoop permiten aos/ás científicos/as de datos axudar ás empresas a reducir custos, aumentar os beneficios, mellorar os produtos, reter clientes e identificar novas oportunidades.

Este curso axuda aos/ás participantes a comprender o que fan os/as data scientists, os problemas que resolven e as ferramentas e técnicas que utilizan. A través de simulacións na clase, os/as participantes aplican os métodos de data science aos retos do mundo real en diferentes industrias e, en última instancia, prepáranse para as funcións de data scientist no campo.

Obxectivos

Ao finalizar a formación, o/a participante saberá utilizar:

- Apache Spark 2 para Data Science e machine learning en fluxos de traballo a escala
- Spark SQL e Dataframes para traballar con datos estruturados
- MLlib, a librería de Spark para machine learning
- PySpark, a API de Python para Spark
- Sparklyr, unha interface de R compatible con dplyr para Spark
- O Cloudera Data Science Workbench (CDSW)
- Outros compoñentes do ecosistema Hadoop: HDFS, Hive, Impala e Hue

Dirixido a

O curso de científico/a de datos está dirixido a enxeñeiros/as de datos e desenvolvedores/as con coñecementos básicos en Data Science e machine learning, así como, para científicos/as de datos que traballaron con Python ou R para pequenos conxuntos de datos nunha única máquina e necesitan escalalo a conxuntos de datos

máis grandes en sistemas distribuídos.

O alumnado debe ter coñecementos básicos en Python ou R e experiencia con análise de datos ou modelos de machine learning. Non se requiren coñecementos en Hadoop ou Spark.

Recomendable ter coñecemento nivel medio comprensión lectora de inglés.

Perfil do docente

O persoal docente ten máis de 5 anos de experiencia en áreas de alta especialización técnica nos ámbitos de aplicación. Dispoñen das certificacións oficiais do fabricante (neste caso Cloudera)para impartir estes cursos.

DURACIÓN	60 horas
PROGRAMA	Programación 2019/20
MATRÍCULA	Gratuíta
METODOLOXÍA	Virtual
TIPO	CURSO
BENEFICIOS	
HORARIO	De luns a venres de 16:30 a 20:30 horas.
PERIODO INSCRICIÓN	27/04/2020 - 06/05/2020
PROBA DE SELECCIÓN	12/05/2020, 16:00
PERIODO DOCENCIA	25/05/2020 - 12/06/2020

LUGAR DE DOCENCIA Edificio localizado na r/Airas Nunes s/n, barrio de Conxo, en

Nº PRAZAS 20 (Mínimo 10)

Temario

Introdución

Data Science

- Que fan os data scientists, ferramentas e procesos que utilizan

Cloudera Data Science Workbench

- Introdución
- Como se utiliza?

Caso de estudo

- Explicación e análise do caso
- Uso de Hue

Apache Spark

- Como traballa Apache Spark e que capacidades nos ofrece
- Que formatos de ficheiros populares pode usar Spark para almacenar datos
- Que linguaxes de programación podes utilizar para traballar con Spark
- Como empezar a utilizar PySpark e sparklyr
- Como comparar PySpark e sparklyr

Machine Learning

- Que é machine learning?
- Algúns conceptos e termos importantes
- Diferentes tipos de algoritmos
- Librerías que se utilizan

Apache Spark MLlib

- Que capacidades de machine learning proporciónanos MLlib
- Como crear, validar e utilizar modelos de machine learning con MLlib

Execución de traballos Apache Spark

- Como un traballo de Spark se compón dunha secuencia de transformacións seguida

dunha acción
- Como Spark utiliza a execución lenta
- Como Spark divide os datos entre as particións
- Como executa Spark operacións limitadas e grandes
- Como Spark executa un traballo en tarefas e fases

Conclusión