Científico/a de datos Big Data Cloudera

Descrición

Os/as científicos/as de datos 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 científicos/as de datos ao permitir a análise de datos interactivos e iterativos a escala.

Aprender como Spark e Hadoop permiten aos/ás científicos/as de datos axudar ás empresas para 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 científicos/as de datos, os problemas que resolven e as ferramentas e técnicas que utilizan. A través de simulacións en clase, as persoas participantes aplican os métodos de data science aos retos do mundo real en diferentes industrias e, en última instancia, prepáranse para as funcións de científicos/as de datos no campo.

Obxectivos

Ao finalizar a formación, o alumnado 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 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.

Os/as estudantes deben 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.

Perfil do docente

Persoas con 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 2018/19
MATRÍCULA	Gratuíta
METODOLOXÍA	Virtual
TIPO	CURSO
BENEFICIOS	
HORARIO	De luns a venres de 16:30 a 20:30 horas.
PERIODO INSCRICIÓN	10/06/2019 - 20/06/2019
PROBA DE SELECCIÓN	26/06/2019, 17:00
PERIODO DOCENCIA	02/09/2019 - 20/09/2019
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 componse 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