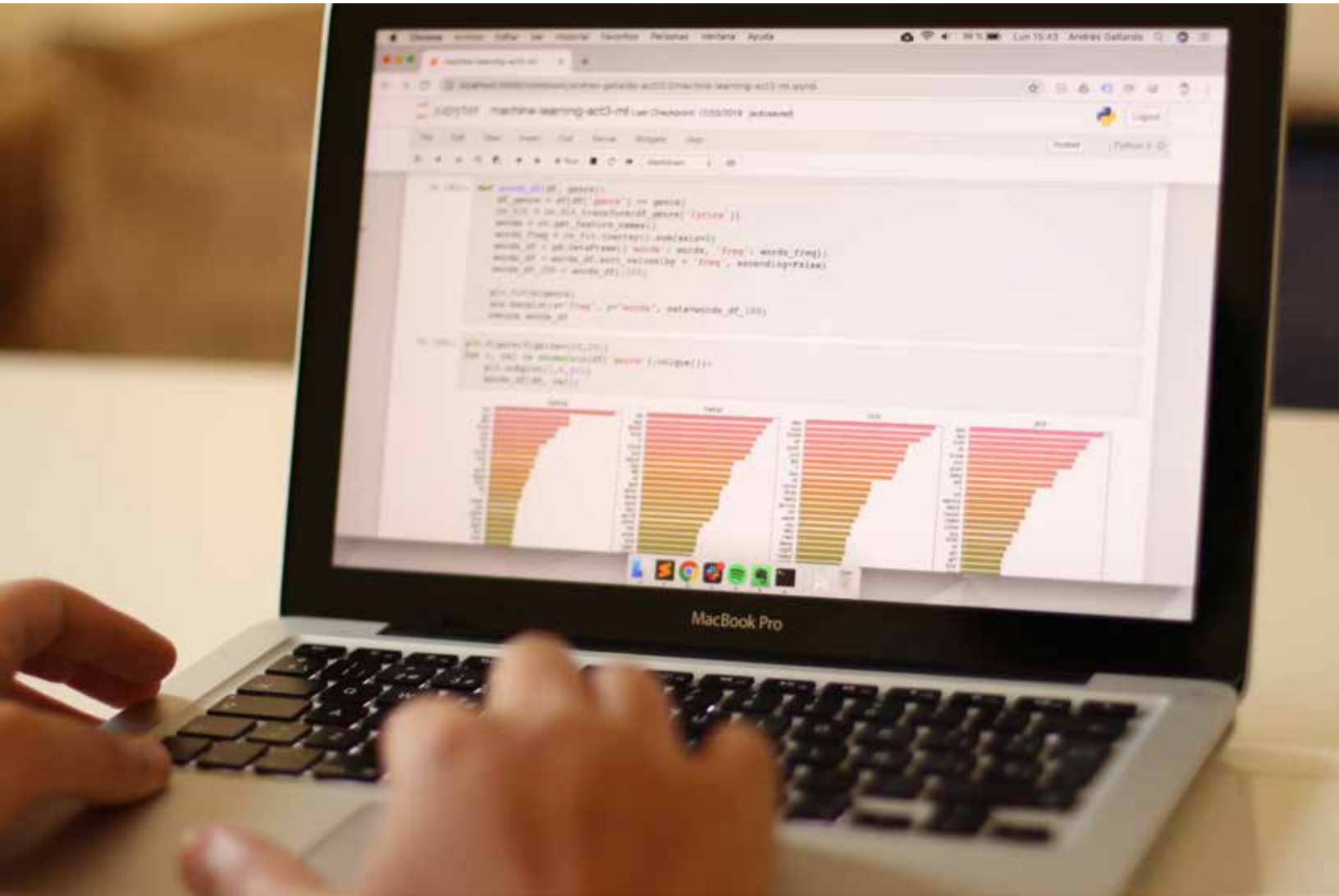
Curso
Big Data

{desafío}
latam_

Descripción del **Programa**

El curso de Big Data te permitirá clasificar los problemas del Big Data según sus características, dimensionar según su volumen. Además podrás escoger las estrategias y herramientas adecuadas para procesar los datos dependiendo de su volumen, utilizando herramientas como Hadoop, Apache Spark y los servicios distribuidos en las nube de Amazon, para analizar grandes flujos de datos sin las limitaciones de un ambiente centralizado.

Este programa es uno de los módulos de la carrera de Data Science que tiene Desafío Latam.



Unidades y **Contenidos**

Unidad 1

Introducción a Big Data

- Conocer los conceptos fundamentales asociados a Big Data así como la importancia del análisis de algoritmo
- Reconocer las características de Big Data.
- Identificar los tipos de soluciones de Big Data.
- Identificar el uso de las funciones map, filter y reduce fundamentales de la programación funcional.
- Conocer las características de una solución de Big Data.
- Comprender los casos de uso de Big Data, tanto exitosos como de fracaso.
- Comprender los distintos paradigmas de soluciones Big Data.
- Conocer los conceptos OLTP y OLAP.
- Conocer los flujos ETL y ELT.
- Conocer los conceptos de Data Lake, Data Warehouse y Data Mart.

Unidad 2

Preparación del ambiente de trabajo

- Conocer los principales beneficios y características de los servicios Cloud asociados al procesamiento masivo de datos.
- Conocer los principales servicios de Amazon Web Services en relación a Big Data.
- Generar buckets de almacenamiento utilizando AWS S3.
- Generar instancias de trabajo utilizando AWS ElasticMapReduce.
- Ejecutar scripts utilizando AWS ElasticMapReduce.
- Implementar conexiones y migraciones de archivos con una instancia de trabajo AWS EMR utilizando ssh y scp.
- Comprender la necesidad de procesar datasets de forma paralela.
- Identificar las diferencias entre procesamiento distribuido y paralelo.
- Implementar un MapReduce primitivo utilizando Python y Bash.

Unidad 3

Introducción a Hadoop

- Conocer los principales componentes del ecosistema Hadoop.
- Conocer cómo funciona la escritura y lectura de archivos en el Hadoop Distributed File System.
- Conocer los términos Maestro y Esclavo en el ecosistema Hadoop.
- Conocer cómo se implementa el manejo de recursos mediante YARN.
- Comprender el uso y cómo se relacionan HDFS, YARN y Hadoop Streaming.
- Utilizar las principales operaciones de Hadoop Distributed File System.
- Implementar MapReduce mediante el jar de Hadoop Streaming en AWS EMR.
- Conocer el rol de Sqoop para transferir datos desde una RDBMS a HDFS.
- Generar configuraciones personalizadas en AWS EMR,
- Implementar comandos de Sqoop para migrar tablas y bases de datos a HDFS.
- Conocer el funcionamiento y objetivos de Hive.
- Generar tablas y queries con Hive.
- Implementar puertos dinámicos para la conexión con interfaces de usuario en AWS EMR.

Unidad 4

Spark I

- Conocer los casos de uso de Spark, así como las diferencias entre MapReduce y Spark.
- Conocer la arquitectura y componentes de Spark.
- Habilitar notebooks de Jupyter desde la instancia AWS EMR.
- Implementar consultas en Spark.
- Utilizar transformaciones y acciones en RDD para extraer resultados.
- Conocer los distintos componentes de la API de Spark.
- Implementar conexiones generalizadas mediante el objeto SparkSession.
- Generar flujos de trabajo con Spark SQL.
- Conocer el objeto y sus componentes de pyspark.sql.dataframe.DataFrame.
- Generar consumos de distintos datos hacia un objeto DataFrame.
- Conocer los principales modos de trabajo con un objeto DataFrame.
- Importar y exportar archivos en formato parquet.

Unidad 5

Spark II

- Conocer los casos de uso de la librería MLlib y ML.
- Conocer los principales tipos de datos asociados para trabajar con Machine Learning en Spark.
- Implementar algoritmos de Machine Learning en Spark.
- Aprender a generar búsqueda de hiperparámetros vía grilla en Spark.
- Generar Pipelines para agilizar el trabajo de una aplicación de Spark basada en Machine Learning.

Duración

- **5 semanas**
- **Sesión online:**
  30 horas (5 sesiones de 6 horas cada una)
- **Sesión presencial:**
  30 horas (10 sesiones de 3 horas cada una)

Requerimientos

Características de tu notebook*

- Empieza (https://empieza.desafiolatam.com)
- Jupyter Notebook
- iPython Kernel
- Hadoop
- Spark
- Amazon Web Services***

** El notebook es por cuenta de todos los participantes: docente, ayudante y alumnos.*
**** Los alumnos requieren un previo registro de cuenta en Amazon Web Services.*

Curso
Big Data

{desafío} latam_

www.desafiolatam.com

/DesafioLatam /DesafioLatam /DesafioLatam /DesafioLatam