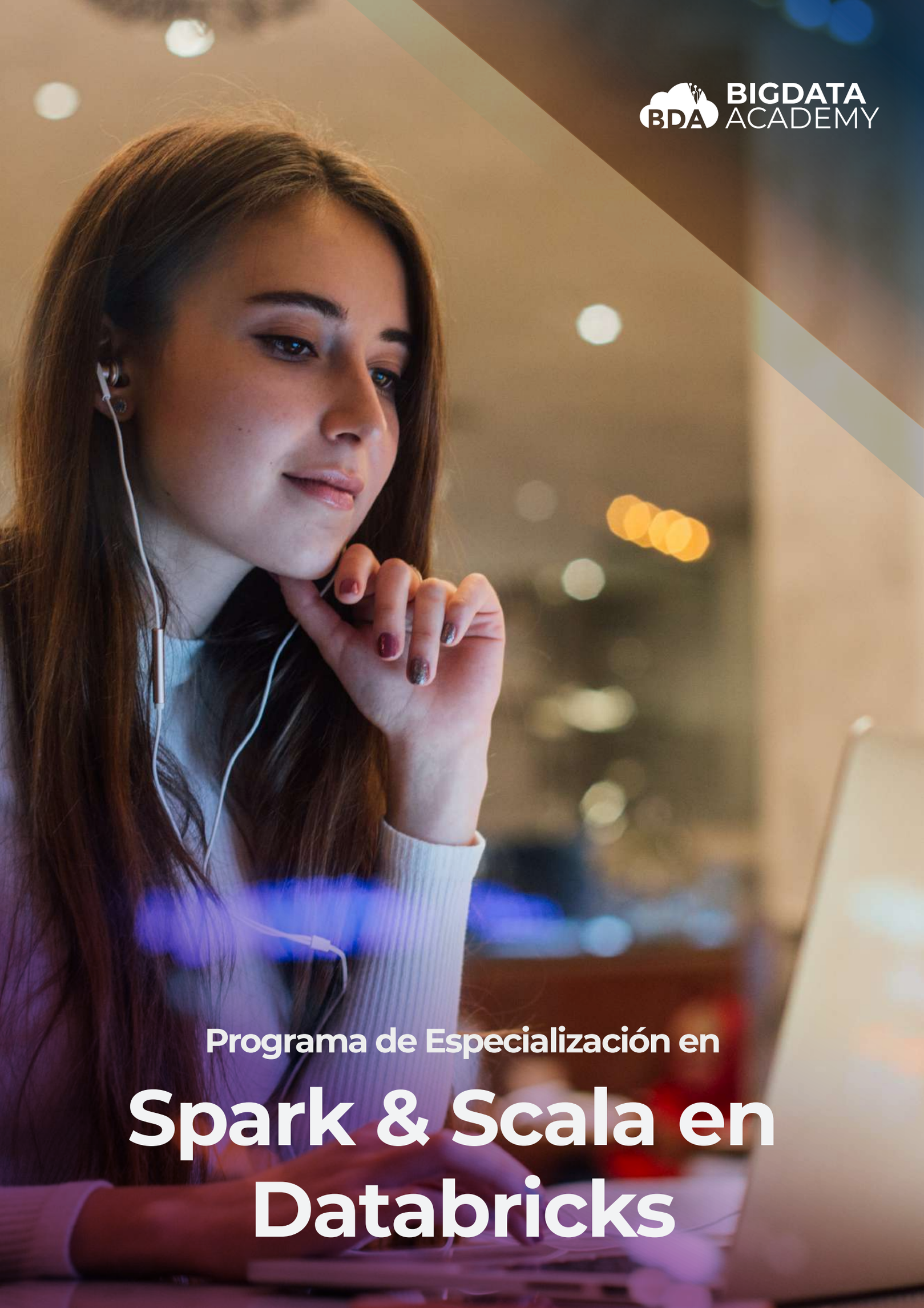


BIGDATA ACADEMY

Programa de Especialización en

Spark & Scala en Databricks

ACERCA DEL CURSO

Con este curso aprenderás a programar con Spark & Scala en un entorno Databricks, desde la sintaxis básica hasta el tuning avanzado. Podrás codificar y procesar datos estructurados y no estructurados por medio de dataframes, udfs, formatos binarios optimizados, procesos en tiempo real, procesos de Machine Learning, procesos de Deep Learning, tuning y mejora de código, puesta en producción, procesamiento y lectura de datos sobre AWS, Azure y GCP, entre otros temas.

Clases:

- 08 Sesiones [28 horas académicas]
- Lunes y Martes de 6:30PM a 9:00PM

Modalidad
- Online con profesor en vivo

Inicio de Clases:
- Lunes 13 de Junio del 2022

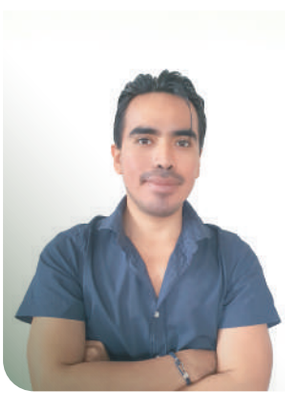
Fin de Clases:
- Martes 05 de Julio del 2022

Egresados con doble certificación:
- Spark & Scala Professional
- Databricks Professional

Perfil del Docente:

Ingeniero de Sistemas de la UNMSM, con especialización en Business Intelligence y Big Data Analytics en ESAN. Con más de 8 años de experiencia nacional e internacional como líder de proyectos, arquitecto de sistemas y big data aplicado a sectores de la banca, Telcos y Gobierno. Conferencista recurrente en ponencias de Big Data. Desempeño como docente universitario de algorítmica y análisis de sistemas en la UNMSM.

Actualmente es responsable de dirigir la estrategia tecnológica de los proyectos de Big Data y Analytics en Everis con el rol de Big Data Architect. Interesado en impulsar la formación profesional en temas que marquen la tendencia tecnológica como el Big data, Blockchain, IoT, IA y Analítica avanzada.



Alonso Melgarejo

- Big Data Architect Senior
- Big Data Architect
- Ing. De Sistemas - UNMSM

alonsoraulmgs@gmail.com
alonsoraulmgs

CURRÍCULA DEL PROGRAMA

MÓDULO 1
INTRODUCCIÓN Y PREPARACIÓN DEL ENTORNO

- Introducción teórica
- Big Data y Spark
- Programación en Drivers vs programación en executors
- Procesamiento In-Memory
- Componentes de Spark
- Variables distribuidas en un clúster: RDD vs DATAFRAME
- Lenguajes de programación para Spark: Scala vs Python vs R
- Despliegue de clúster Spark
- Configuración a repositorio de datos

MÓDULO 2
PROCESAMIENTO ESTRUCTURADO CON SPARK SQL

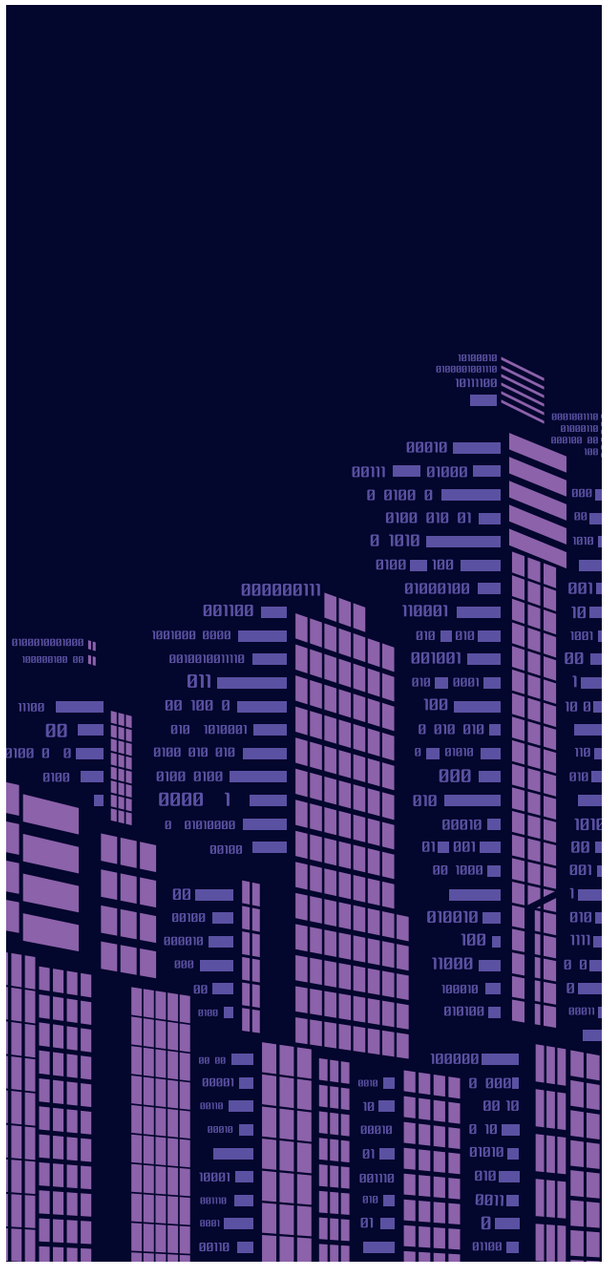
- Definición del “SparkSession” para procesos batch
- Lectura de archivos como tablas estructuradas
- Definición de schemas
- Transformations en SQL
- Creación de vistas temporales
- Almacenamiento en formatos binarios de rápido procesamiento: Parquet y Avro

MÓDULO 3
PROCESAMIENTO ESTRUCTURADO Y SEMI-ESTRUCTURADO CON DATAFRAMES

- Programación funcional
- Dataframes para datos estructurados y semi-estructurados
- Transformations y operations en dataframes
- Creación de funciones personalizadas con UDFs
- UDFs con multi-parámetros
- Almacenamiento en formatos binarios de rápido procesamiento: Parquet y Avro

MÓDULO 4
TUNING Y PATRONES DE DISEÑO PARA BATCH

- Patrón de checkpoint para evitar colapso de memoria RAM
- Patrón de caché para reducir tiempos de procesamiento
- Patrón delta-lake para actualizaciones
- Aumentando el nivel de paralelización con executors
- Gestión de particiones de los dataframes
- Monitoreo con Spark UI



CURRÍCULA DEL PROGRAMA

MÓDULO 5

CONEXIÓN A AWS, AZURE Y GCP

- Databricks como entorno Multi-Cloud para procesamiento en Spark
- Integración, conexión y lectura de datos desde el S3 de AWS
- Integración, conexión y lectura de datos desde el Blob Storage de Azure
- Integración, conexión y lectura de datos desde el Cloud Storage de GCP

MÓDULO 6

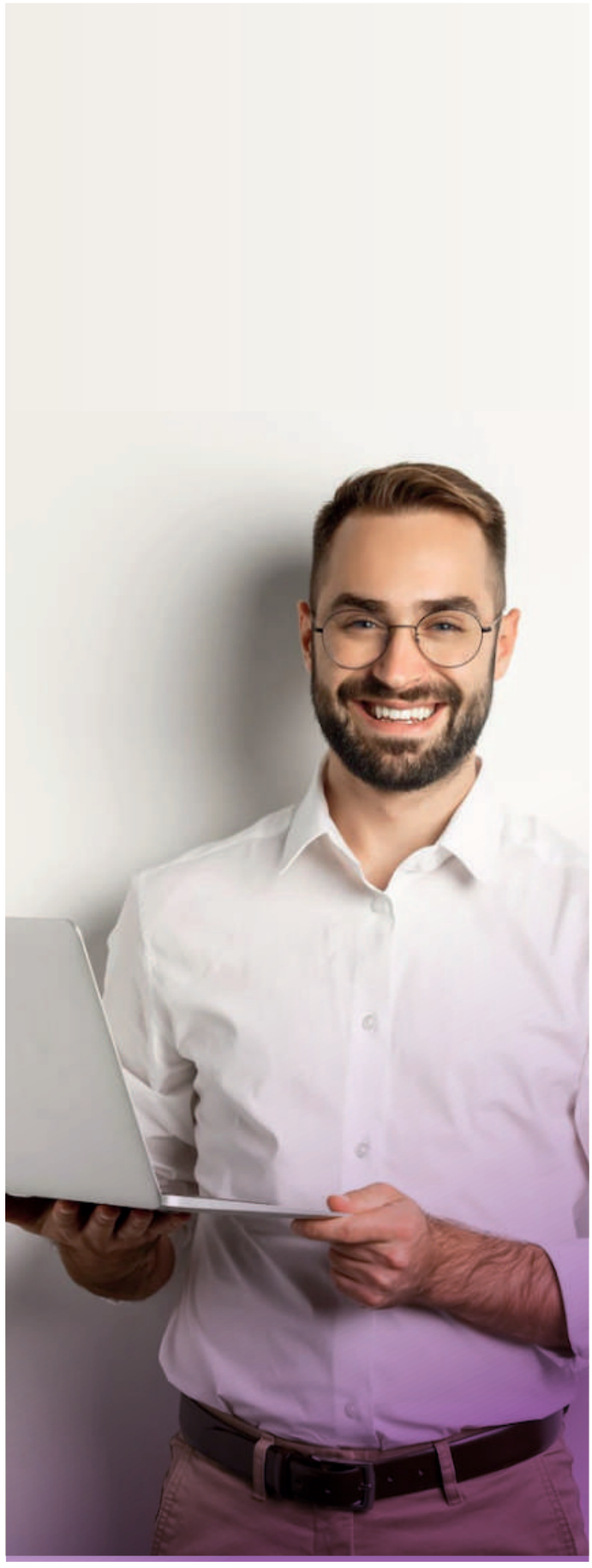
SPARK STREAMING PARA PROCESAMIENTO REAL-TIME

- Arquitectura real-time y de storm data
- Kafka como interfaz estándar de procesamiento
- Patrón micro-batch para optimización de ahorro de tiempo en CPU
- Patrón de diseño producer: read, format & write
- Patrón de diseño consumer: read, format, enrichment & process
- Integración a Kinesis de AWS
- Integración a EventHub de Azure
- Integración a Pub/Sub de GCP

MÓDULO 7

PROCESAMIENTO ANALÍTICO CON SPARK MACHINE LEARNING LIBRARY

- Analítica y Machine Learning
- Analítica descriptiva y predictiva
- Clusterización con KMeans
- Deep Learning & Redes Neuronales
- Almacenamiento y despliegue de modelos
- Uso de modelos analíticos en producción



BENEFICIOS DEL CURSO

- Aprende todo el mix tecnológico: batch, real-time y machine learning.
- Aprenderás a conectarte y procesar en las tres nubes: AWS, Azure y GCP.
- Trabajarás con la última versión de Spark: Spark 3.
- Integrarás todo lo aprendido dentro de una arquitectura Batch, Streaming y de Machine Learning que se desplegará en la nube.
- Al terminar la clase el profesor enviará la grabación de video para que puedas repasar.

REQUISITOS

- Conocimientos básicos en SQL..

MATRÍCULA Y PROCESOS DE PAGO

1 Matrícula automática por la Web

Ingresa a **www.bigdataacademy.org** y haz clic en la página de **"cursos"**. Selecciona el Programa de especialización de preferencia e inglresa tus datos. Realiza el pago con cualquier tarjeta de crédito/ débito.

VISA mastercard Diners Club INTERNATIONAL AMERICAN EXPRESS

Al finalizar recibirás un correo de confirmación.

2 Depósito / Transferencia a BCP

BIG DATA ACADEMY PERU SAC
Ruc: 20603117655
Cta Ahorros BCP Soles: N° 193-2514329-0-61
Cta Interbancaria BCP Soles: N°
00219300251432906114

3 Envío de comprobante

Enviar comprobante de pago a **info@bigdataacademy.org** indicando datos del alumno.

Asunto: **"Matrícula Spark".**
Nombres y Apellidos: .....
DNI: .....
Celular: .....
Correo Electrónico: .....
Puesto y lugar de trabajo: .....

4 Confirmación

Confirmaremos su inscripción a la brevedad por el mismo medio o vía telefónica.

Contacto

info@bigdataacademy.org
Cel:943516891
f/bigdataacademyperu/
http://bigdataacademy.org