CTIC UNI

UNIVERSIDAD NACIONAL DE INGENIERÍA

OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

Programa de Iniciación Tecnológica PIT - 2026

CERTIFICADO

Otorgado a

DAVID DANIEL ROMERO SUXE

Por haber aprobado el curso de

CIENCIA DE DATOS 2: BIG DATA

Realizado del 23 de mayo al 04 de julio del presente, con una duración de 18 horas.

Lima, julio de 2026

Mag. Ing. Rubén Arturo Borja Rosales
JEFE OTI - UNI

N° Certificado: 018 - 0028089

FIRMA DIGITAL

Firmado digitalmente por: BORJA ROSALES Ruben Arturo FAU 20168004359 hard
Motivo: En señal de conformidad
Fecha: 25/07/2026 11:54:45-0500

UNIVERSIDAD NACIONAL DE INGENIERÍA

OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

CTIC UNI

Temario del Curso:

CIENCIA DE DATOS 2: BIG DATA

- Introducción y fundamentos
- Paradigma Big Data y el rol de Spark en la era de los LLMs
- Arquitectura de Spark (Driver, Executors, Cluster Manager)
- DataFrames vs. RDDs y casos de uso
- Transformaciones distribuidas con PYSparK
- Operaciones básicas: Select, Filter, Groupby, Joins, UDFs
- Control de esquemas y tipos de datos
- Particionado y Spark.sql.shuffle.partitions
- Procesamiento de texto a escala
- Limpieza robusta: Unicode, HTML, espacios, Lowercasing
- Tokenización y Stopwords distribuidas
- Detección de idioma, spam y toxicidad básica
- SparkSQL y optimización
- SparkSQL: consultas declarativas y Catalyst Optimizer
- Delta Lake como estándar (ACID, schema evolution, MERGE INTO)
- Técnicas de optimización:AQE, Broadcast Joins, Salting Anti-skew
- Curación y preparación LLM-READY
- Curación final de datasets textuales (Filtrado, normalización, estratificación)
- Validación de calidad de datos con Great Expectations/Deequ
- Dataset Card y metadatos
- Pipeline E2E y despliegue
- Orquestación simple (Databricks Jobs / Prefect / Airflow)
- Reproducibilidad y linaje de datos (Bronze → Silver → Gold)
- Structured Streaming (opcional): lectura incremental con checkpointing

Nota final: **20 (Veinte)**

N° Certificado: 018 - 0028089