CTIC UNI

UNIVERSIDAD NACIONAL DE INGENIERÍA

OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

Programa de Iniciación Tecnológica PIT - 2025

CERTIFICADO

Otorgado a

JEAN CARLOS ZUBIATE HUAMANI

Por haber aprobado el curso de

CIENCIA DE DATOS 2: BIG DATA + IA

Realizado del 11 de noviembre al 27 de noviembre del presente, con una duración de 18 horas.

Lima, noviembre de 2025

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

N° Certificado: 017 - 0132445

FIRMA DIGITAL

Firmado digitalmente por: BORJA ROSALES Ruben Arturo FAU 20168004359 hard
Motivo: En señal de conformidad
Fecha: 23/12/2025 23:29:53-0500

UNIVERSIDAD NACIONAL DE INGENIERÍA

OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

CTIC UNI

Temario del Curso:

CIENCIA DE DATOS 2: BIG DATA + IA

- 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.
- Configuración del entorno (Repo estructurado, logging, testing básico).
- Transformaciones distribuidas con Pyspark.
- Operaciones básicas: Select, filter, groupBy, joins, UDFs.
- Control de esquemas y tipos de datos.
- Particionado y spark.sql.shuffle.partitions.
- Validaciones ligeras con tests en DataFrames.
- Procesamiento de texto a escala.
- Limpieza robusta: Unicode, HTML, espacios, lowercasing.
- Tokenización y stopwords distribuidas
- Detección de idioma, spam y toxicidad básica.
- Deduplicación con hashing/LSH.
- Enmascaramiento/anónimo de PII (Emails, teléfonos, IDs).
- 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.
- Persistencia y caching; particionamiento y Z-Order en Delta.
- 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.
- Export a formatos estándar (Parquet, JSONL, HuggingFace Datasets).
- 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.
- Integración con MLflow para registrar métricas de dataset.
- Publicación en HF Datasets o repositorio interno.

Nota final: **20 (Veinte)**

N° Certificado: 017 - 0132445