CTIC UNI

# UNIVERSIDAD NACIONAL DE INGENIERÍA

## OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

### Programa de Iniciación Tecnológica PIT - 2025

# *CERTIFICADO*

**Otorgado a**

## YUSTIN HERNAN GALOPINO LOPEZ

---

**Por haber aprobado el curso de**

**CIENCIA DE DATOS 2: BIG DATA**

**Realizado del 02 de agosto al 14 de agosto del presente, con una duración de 16 horas.**

Lima, setiembre de 2025

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

**N° Certificado: 017 - 0075342**

FIRMA DIGITAL

Firmado digitalmente por: BORJA ROSALES Ruben Arturo FAU 20169004359 hard
Motivo: En señal de conformidad
Fecha: 12/09/2025 17:13:03-0500

# UNIVERSIDAD NACIONAL DE INGENIERÍA

## OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

CTIC UNI

**Temario del Curso:**

### CIENCIA DE DATOS 2: BIG DATA

- Introducción a Big Data en la era de la IA.
- Qué es Big Data y cómo lo utilizan los LLMs (GPT, LLaMA, Claude, etc.).
- El ciclo de vida de los datos en proyectos de IA.
- ¿Dónde entra Spark? Procesamiento escalable antes de modelar.
- Fundamentos del procesamiento distribuido con PySpark.
- ¿Por qué es necesario distribuir el procesamiento?
- RDD vs DataFrames: Diferencias clave y casos de uso.
- Arquitectura básica de Spark: Driver, Executors, Cluster Manager.
- Limpieza y transformación de datos a gran escala.
- Técnicas comunes de preprocesamiento para textos y metadatos.
- Transformaciones con PySpark: Select, Filter, groupBy, Joins.
- Introducción a funciones UDF (User Defined Functions).
- Optimización y consultas con SparkSQL.
- ¿Por qué importa optimizar antes de entrenar modelos?
- SparkSQL: Consultas SQL en datos distribuidos.
- Catalyst Optimizer y almacenamiento columnar (Parquet).
- Datos textuales para LLMs: Preparación y curación.
- ¿Qué tipo de datos usan los LLMs? Estructura típica de los datasets.
- Tokenización, filtrado y limpieza profunda (Remover duplicados, lenguaje ofensivo, etc.).
- Pre-procesamiento para embeddings, entrenamiento o fine-tuning.
- Proyecto Final: Pipeline de datos para LLM.
- Simulación de un Pipeline completo: Desde datos brutos hasta Dataset Usable.
- Integración con otras herramientas: Hugging Face Datasets, LangChain, o LLM-ready formats.
- Buenas prácticas para escalar a producción o entrenamiento real.

**Nota final:** **20 (Veinte)**

**N° Certificado: 017 - 0075342**