CTIC UNI

UNIVERSIDAD NACIONAL DE INGENIERÍA

OFICINA DE TECNOLOGÍAS DE LA INFORMACIÓN

Programa de Iniciación Tecnológica PIT - 2025

CERTIFICADO

Otorgado a

WALTER URBANO GARCIA MACAVILCA

Por haber aprobado el curso de

CIENCIA DE DATOS 2: BIG DATA

Realizado del 23 de mayo al 09 de junio del presente, con una duración de 16 horas.

Lima, julio de 2025

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

N° Certificado: 017 - 0061570

FIRMA DIGITAL

Firmado digitalmente por: BORJA ROSALES Ruben Arturo FAU 20168004359 hard
Motivo: En señal de conformidad
Fecha: 08/07/2025 10:52:20-0500

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.
- Práctica: Cargar y explorar un dataset textual grande en PySpark.
- 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 - 0061570