



Extensión y
Proyección
Universitaria

Facultad de Ingeniería
y Arquitectura

Curso

Ingeniería de Datos con Databricks y Spark



Horario: Lunes, miércoles y viernes
de 19:00 a 21:00 h.

Duración: 24 horas

Modalidad: Virtual

Inversión:

Público general: S/. 600.00

Alumnos, egresados y

docentes USMP S/. 500.00

Descripción del curso

En la era del Big Data, las organizaciones necesitan profesionales capacitados en el procesamiento y análisis de grandes volúmenes de datos de manera eficiente y escalable. Databricks, junto con Apache Spark, se ha convertido en la plataforma líder para ingeniería de datos y análisis avanzado.

Este curso proporciona las competencias prácticas necesarias para diseñar, implementar y optimizar pipelines de datos modernos utilizando Databricks y Apache Spark, preparando a los participantes para enfrentar los desafíos del procesamiento de datos a gran escala.

Objetivos

- Proporcionar conocimientos teóricos y prácticos sobre la arquitectura y funcionamiento de Apache Spark y la plataforma Databricks.
- Desarrollar habilidades para el diseño e implementación de pipelines ETL/ELT robustos y escalables.
- Capacitar en la optimización del rendimiento de aplicaciones Spark y el uso eficiente de recursos en Databricks.
- Transmitir las mejores prácticas en ingeniería de datos moderna y arquitecturas de lakehouse.

Dirigido a

Ingenieros de datos, analistas de datos, científicos de datos, desarrolladores de software interesados en procesamiento de Big Data, arquitectos de datos, profesionales de TI que buscan especializarse en tecnologías de datos modernas, y en general, profesionales técnicos que trabajan o desean trabajar con procesamiento distribuido de datos a gran escala.

Metodología

- Virtual, mediante la plataforma Zoom.
- Clases teórico-prácticas con ejercicios hands-on en Databricks Community Edition.
- La metodología del curso está orientada al logro de los objetivos mediante la práctica directa con notebooks, datasets reales y casos de uso empresariales

Requisitos

Conocimiento básico de SQL.

Conocimiento básico de Python.

Requisitos técnicos

Computadores 2 núcleos y 8 GB de ram.

TABLA DE CONTENIDO

Módulo 1

INTRODUCCIÓN A BIG DATA Y ARQUITECTURA SPARK

- Conceptos fundamentales de Big Data y procesamiento distribuido.
- Arquitectura de Apache Spark: componentes, modelo de ejecución y ecosistema.
- Databricks: plataforma, workspace, clusters y notebooks.
- Configuración del entorno de desarrollo y primeros pasos con Databricks Community Edition.
- Conceptos de RDD, DataFrames y Datasets.

Módulo 2

PROCESAMIENTO DE DATOS CON SPARK DATAFRAMES

- Fundamentos de Spark DataFrames API.
- Lectura y escritura de datos en múltiples formatos (CSV, JSON, Parquet, Delta).
- Operaciones de transformación: select, filter, groupBy, join, window functions.
- Manejo de datos faltantes y limpieza de datos.
- Uso de Spark SQL para análisis de datos.
- Integración con catálogos de datos (Unity Catalog).

Módulo 3

DISEÑO E IMPLEMENTACIÓN DE PIPELINES ETL/ELT

- Arquitecturas modernas de datos: Data Lake, Data Warehouse y Lakehouse.
- Diseño de pipelines ETL con medallion architecture (Bronze, Silver, Gold).
- Delta Lake: fundamentos, operaciones ACID y versionamiento de datos.
- Procesamiento batch vs streaming: casos de uso y mejores prácticas.
- Orquestación de workflows con Databricks Jobs y Delta Live Tables.
- Gestión de dependencias y parametrización de pipelines.

Módulo 4

OPTIMIZACIÓN Y MEJORES PRÁCTICAS

- Optimización del rendimiento en Spark: particionamiento, bucketing y caching.
- Análisis del Spark UI y métricas de ejecución.
- Gestión eficiente de recursos: configuración de clusters y auto-scaling.
- Manejo de data skew y optimización de joins.
- Mejores prácticas en control de versiones, testing y CI/CD para pipelines de datos.
- Seguridad, gobernanza de datos y compliance en Databricks.
- Monitoreo, logging y troubleshooting de aplicaciones Spark.

Docente: José Alonso Medina Donayre

Profesional con más de 8 años de experiencia diseñando y desarrollando soluciones de datos. Cuenta con certificaciones profesionales y associate en Databricks y Microsoft Azure. Se desempeña además como instructor en programas especializados de datos en reconocidas universidades y academias, y comparte conocimiento activamente mediante contenido educativo en LinkedIn, Medium y YouTube.

Inicio de clases: 06 de abril
Fin de clases: 04 de mayo

Inscríbete dando clic aquí



Se entregará certificado y material del curso.

*Cronograma y plana docente sujeta a variación.



Informes:

Oficina de Extensión y Proyección Universitaria
epu_fia@usmp.pe
(511) 208 6000 anexo: 1130
WhatsApp: 967 304 157

Av. La Fontana 1250. Urb. Santa Patricia, La Molina