



Cursos AMAT

 **NUEVO Curso de: //**

**Procesamiento de datos
a gran escala con**

PySpark APACHE  TM



OBJETIVO

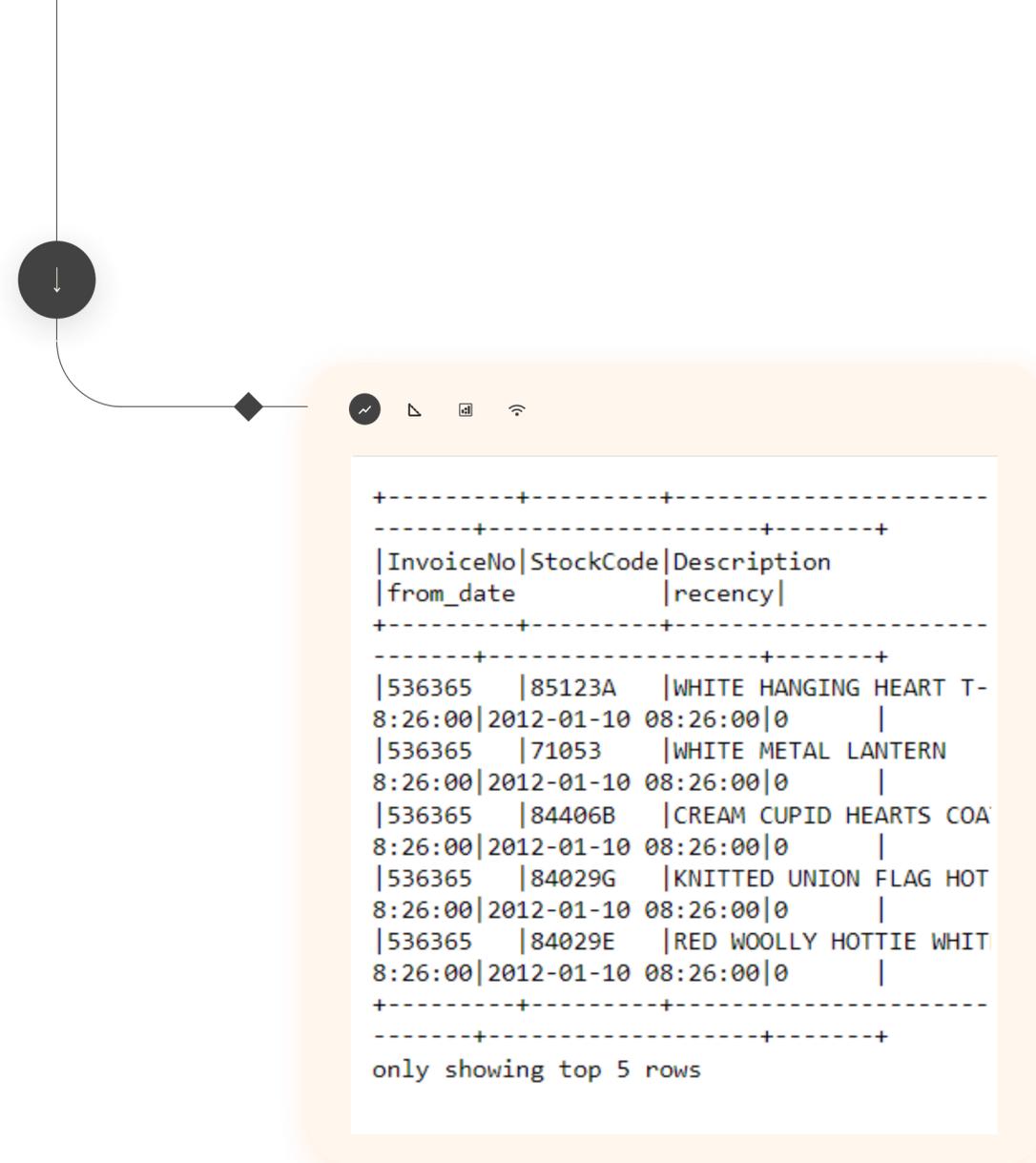
Proporcionar a los participantes mediante módulos prácticos las habilidades y conocimientos necesarios para utilizar PySpark en el análisis de datos, incluyendo su manipulación, aplicación de SQL, machine learning y más.

¿Sabías que...?

Apache Spark es un motor de análisis unificado de código abierto para el procesamiento de datos a gran escala.

Creado en AMPLab de UC Berkeley en 2009.

Rápida adopción y contribución por parte de la comunidad de código abierto

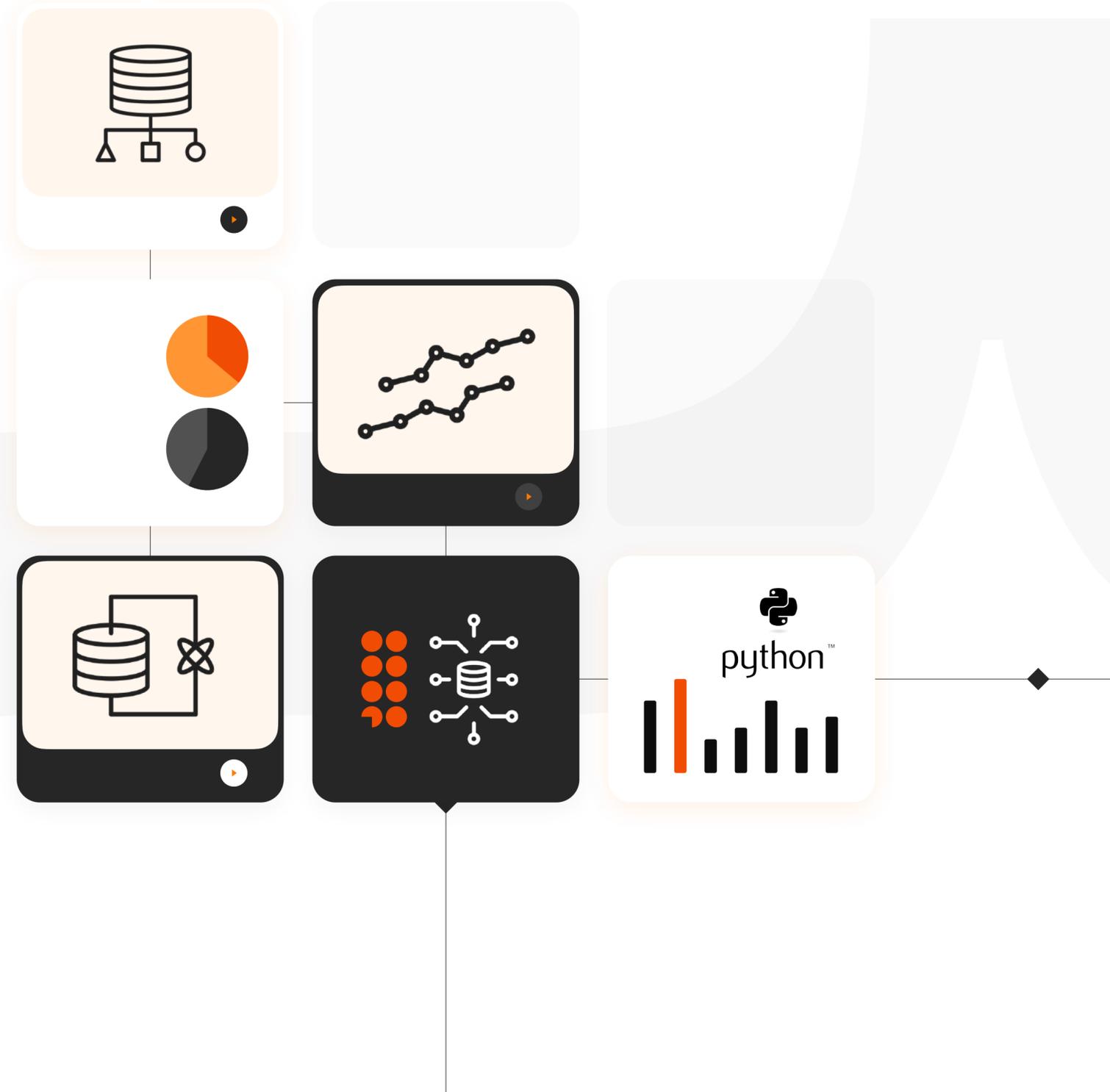


¿A quién va dirigido?

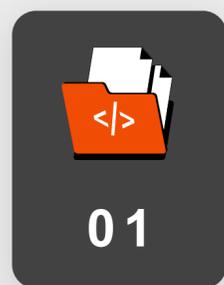
Profesionales con nociones básicas de Python que desean profundizar en el uso de PySpark para el análisis y procesamiento de grandes volúmenes de datos.

Al finalizar este curso:

Tendrás una base sólida en el uso de PySpark para el análisis de datos, abarcando desde operaciones básicas con RDDs y DataFrames hasta técnicas avanzadas como el uso de GraphFrames.



Temario del curso



01

Introducción a PySpark

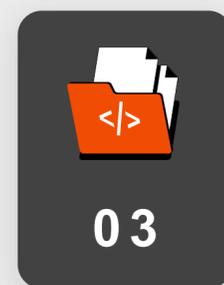
1. Introducción a Big Data y Apache Spark
2. Instalación y configuración de PySpark
3. Introducción a Apache Spark y componentes principales
4. Configuración del entorno de trabajo
5. Primeros pasos con SparkContext y SparkSession



02

Transformaciones y Acciones en RDDs

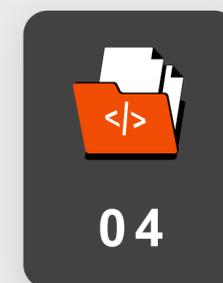
1. Entendiendo los RDDs (Resilient Distributed Datasets)
2. Acciones y Transformaciones básicas
3. Operaciones comunes con RDDs (map, filter, reduce)
4. Práctica: Ejercicios básicos con RDDs
5. Optimización y conceptos clave de los RDDs



03

Introducción a DataFrames y Datasets

1. Concepto de DataFrames y Datasets
2. Creación y manipulación de DataFrames
3. Interacción entre RDDs y DataFrames
4. Operaciones comunes: select, filter, groupBy
5. Práctica: Operaciones básicas con DataFrames



04

Lectura y Escritura de Datos

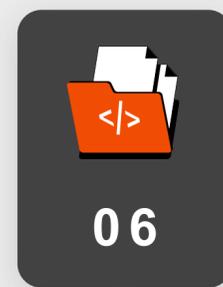
1. Introducción a las fuentes de datos soportadas
2. Lectura de datos desde CSV, JSON, Parquet
3. Escritura de datos a diferentes formatos
4. Manejo de esquemas y inferring
5. Práctica: Ejercicios de lectura y escritura

Temario del curso



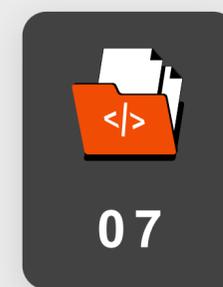
SQL con PySpark

1. Introducción a Spark SQL
2. Uso de SQLContext o SparkSession para consultas SQL
3. Registro de DataFrames como tablas temporales
4. Ejecución de consultas SQL en Spark
5. Práctica: Ejercicios de consultas SQL



Manejo de Datos y Funciones de Usuario (UDF)

1. Transformaciones avanzadas de DataFrames
2. Aplicación de funciones de usuario (UDFs)
3. Creación y registro de UDFs en PySpark
4. Práctica: Uso de UDFs en transformaciones complejas



Machine Learning con PySpark

1. Introducción a MLlib
2. Construcción y entrenamiento de modelos de Machine Learning
3. Evaluación y ajuste de modelos
4. Práctica: Casos de uso de ML con PySpark



M.C. CORINA
CEREZO SILVA

PROFESORES QUE IMPARTEN:

Formación Académica:

- Maestría en Ciencias Matemáticas, UNAM IIMAS.
- Especialidad en Estadística Aplicada, UNAM IIMAS.
- Licenciatura en Matemáticas Aplicadas y Computación, UNAM FES ACATLÁN.

Experiencia Profesional:

- Consultor Sr. Analítica de Clientes, Banco Azteca, Analista en el área de Sistema de Pagos.
- Docente de Estadística, UNAM IIMAS, en la Especialización en Estadística Aplicada.
- Docente de Matemáticas, Universidad Cristóbal Colón, de probabilidad y estadística en la licenciatura de Actuaría.

Software

Programación en:

- ▶ Excel (VBA).
- ▶ Tableau.
- ▶ Sheets.

Paquetería Estadística:

- ▶ Python.
- ▶ R.



M.C. GERARDO
GONZALEZ

Software

Programación en:

- ▶ Java
- ▶ Python
- ▶ C++

Frameworks:

- ▶ Spring Boot
- ▶ Flask

Tecnologías:

- ▶ Kafka
- ▶ ElasticSearch
- ▶ Solr
- ▶ Spark

Herramientas:

- ▶ SQL
- ▶ Git

PROFESORES QUE IMPARTEN:

Formación Académica:

- Maestría en Ciencias de la Computación, Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS).
- Licenciatura en Matemáticas Aplicadas y Computación, Universidad Nacional Autónoma de México

Experiencia Profesional:

- Software Engineer II, Pinterest, Ciudad de México
- Integraciones de servicios relacionados con LLMs.
- Preprocesamiento de datos en Spark para su uso posterior por tecnologías LLMs.
- Creación de microservicios RESTful escalables usando Python.
- Sr. Application Developer, Oracle.
- Creación de microservicios RESTful robustos y escalables para clientes y socios.
- Desarrollo de servicios en la nube con tecnologías serverless en OCI-Oracle.
- Diseño de arquitecturas robustas para microservicios.

```
Cmd 16
Filter like and rlike

1 data2 = [(2,"Michael Rose"),(3,"Robert Williams"),
2         |(4,"Rames Rose"),(5,"Rames rose")
3         ]
4 df2 = spark.createDataFrame(data = data2, schema = ["i
5
6 # like - SQL LIKE pattern
7 df2.filter(df2.name.like("%rose%")).show()
8
9 # rlike - SQL RLIKE pattern (LIKE with Regex)
10 #This check case insensitive
11 df2.filter(df2.name.rlike("(?i)^*rose$")).show()
```

▶ (6) Spark Jobs

▶ df2: pyspark.sql.dataframe.DataFrame = [id: long, name: string]

id	name
5	Rames rose

id	name
2	Michael Rose
4	Rames Rose
5	Rames rose



MODALIDAD

100% Live Streaming

- Con uso de la mejor plataforma a nivel mundial para transmisión en vivo.
- Clases totalmente en vivo.
- Preguntas al instructor en tiempo real.
- Alta calidad en audio y video.
- Conéctate desde tablet, celular o laptop.
- Sólo requieres de una conexión a internet.



INVERSIÓN

Estudiantes y ex alumnos AMAT

Válido al 17 de septiembre 2024

\$4,699

A partir del 18 de septiembre

\$4,999

↓ Público en general

Válido al 17 de septiembre 2024

\$5,199

A partir del 18 de septiembre

\$5,499



→ **A CONSIDERAR...**



- En caso de requerir factura, favor de solicitarla al momento de la inscripción ya que solo se podrá efectuar dentro del mes en que se realizó el pago del curso.
- Si existe cancelación del curso por parte de AMAT, a los participantes que hayan realizado alguna aportación, le será devuelta su inversión, o bien, se les hará válida la aportación para otros cursos.
- Si el alumno desea realizar la cancelación de inscripción, la penalización será equivalente a un 50% del monto que haya depositado. Una vez iniciado el curso la penalización por cancelación de curso será del 90% del valor depositado hasta ese momento y no podrá ser utilizado para el pago o apartado de otro curso.

▶ NUEVO Curso de: //

**Procesamiento de datos
a gran escala con**
PySpark APACHE 

**¡Contáctanos
y participa!**

 info@amatinfo.com

 +52 55 55 44 07 51

 amat.mx

