

# Data Engineer

## CURSO

**Duración: 40 horas**  
**Constancia y diploma con valor curricular**

### Módulo 1:

Hadoop HDFS  
Sistemas distribuidos y YARN  
Sqoop import / export  
Masterización de datos

### Modulo 3

Scala  
Jupyter  
Spark RDD  
Funciones y recursión Scala  
Transformaciones

### Módulo 5

SparkSQL y Hive  
Spark y MySQL via JDBC  
Inferencia de esquemas  
ORC vs Parquet

### Módulo 2:

Near real time  
Kafka  
Streaming  
Streaming esquematizado

### Módulo 4

Spark SQL  
Manejo de datos SQL  
Transformación y acciones  
Tablas temporales y funciones windows  
Funciones definidas en Spark SQL

### Módulo 6

Spark Streaming  
Manejo de datos en tiempo real  
RDD en streaming  
Arquitectura Lambda - Near Real time

Tecnologías que se ven en el curso:



[hola@greenshark.com](mailto:hola@greenshark.com)

55-23-12-23 | (55)-32-23-12

Calle numero #50 Int. 45 Colonia Cuauhtemoc, CDMX, México 07929.