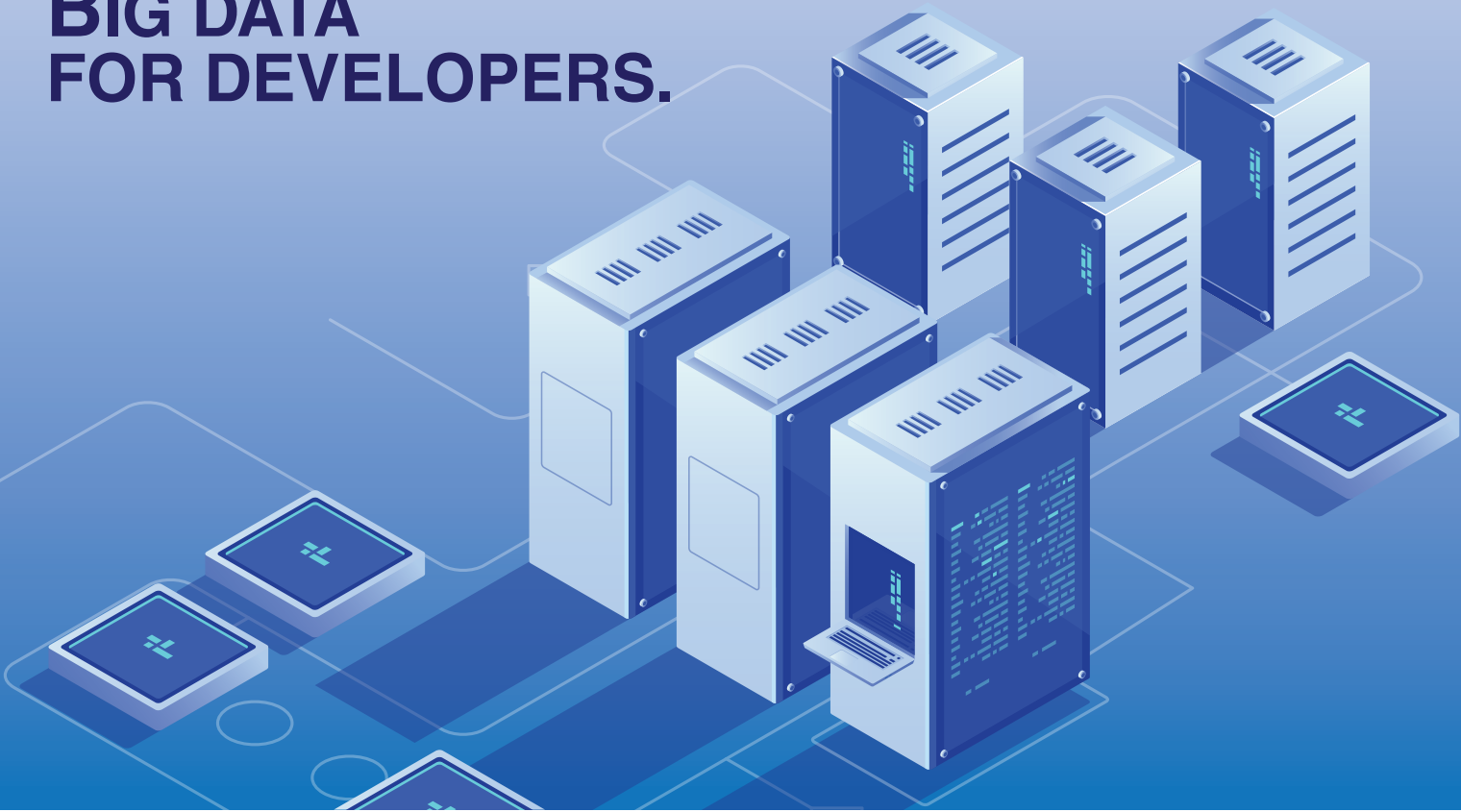




**BIG DATA
FOR DEVELOPERS.**



1. MOTIVACIÓN

Muchos de los cálculos que se realizan en operaciones ETL son conceptualmente fáciles de entender.

Sin embargo, cuando los datos de entrada son grandes, los datos y cálculos deben distribuirse entre varias máquinas para finalizar en un tiempo razonable.

Paralelizar el cómputo, distribuir los datos y manejar las fallas pueden generar un código muy complejo y difícil de mantener.

Este curso se enfoca en ayudar a los desarrolladores a crear programas capaces de escalar, sin importar el volumen de datos, al poder utilizar todo un clúster de máquinas.

2. OBJETIVOS

Se busca que al terminar el curso los participantes entiendan:

- Los principales retos que los grandes volúmenes de información presentan: Volumen, variabilidad y velocidad.
- La historia de cómo se llegó a lo llamado hoy **Big Data**.
- El ecosistema de Apache Hadoop de manera general.
- El sistema de almacenamiento de archivos de Apache Hadoop (HDFS).
- La administración de contenedores en clústers de Hadoop con Apache YARN.
- Introducción a Apache Spark con Scala: RDDs.
- Interfaces de alto nivel de Apache Spark: Dataset, DataFrames, SparkSQL.
- Diseño de aplicaciones para ambientes productivos.



3. TEMARIO

1. Introducción

1.1 *Scala*

1.2 *Introducción a Big Data*

1.2.1 - ¿Por qué?

1.2.2 - ¿Qué es?

1.2.3 - ¿Cómo funciona?

1.3 *Hadoop*

1.3.1 - Historia

1.3.2 - Ecosistema de aplicaciones

1.3.3 - Beneficios

1.3.4 - Distribuidores

1.3.5 - Arquitectura

1.3.6 - Apache Ambari

1.4 *Apache YARN*

1.4.1 - Arquitectura de YARN

1.4.2 - Parametrización de YARN

1.5 *Instalación de ambientes Hadoop con Apache Ambari*

2. Apache Spark

2.1 *Definición y conceptos básicos*

2.2 *Interfaces de alto nivel*

2.2.1 - DataFrames

2.2.2 - DataSet

2.2.3 - SparkSQL

2.3 *Diseño de código para ambientes productivos*

3. Apache Hbase

3.1 *Casos de Uso*

3.2 *Arquitectura*

3.1 *CRUD*

4. FORMATO

- Curso presencial.
- Se busca que sea 60% práctica y 40% teoría.
- El instructor guiará a los alumnos con sesiones de programación en vivo.
- Dos ayudantes guiarán a los participantes que tengan dudas.
- Se creará un clúster con el ambiente Hadoop con máquinas virtuales utilizando VirtualBox, Vagrant y Ambari.

- Máximo 20 alumnos.
- Cada alumno debe contar con una computadora.
- Se dará a cada participante una USB con el material necesario para el curso.
- Es recomendable tener experiencia con la terminal de linux.

5. REQUISITOS



6. INSTRUCTORES

Nuestros instructores cuentan con las siguientes habilidades:

- ✓ **Contribuyen** a proyectos de innovación en diversos temas de la industria.
- ✓ **Asesoran y guían** en diversas implementaciones de **Big Data**.
- ✓ Tienen conocimiento de desarrollo en diversos lenguajes de programación como **Java, Python, Scala**.
- ✓ Cuentan con **experiencia** en **Data Science y Machine Learning**.
- ✓ Manejan ampliamente diversas dinámicas de integración.

