Programa de formación
MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de
INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Módulo 3
Big Data

Contenido inicial

Facultad de INGENIERÍA
UNIVERSIDAD NACIONAL DE COLOMBIA

¡Le damos la bienvenida al tercer módulo del programa de formación MLDS: *Big Data!*

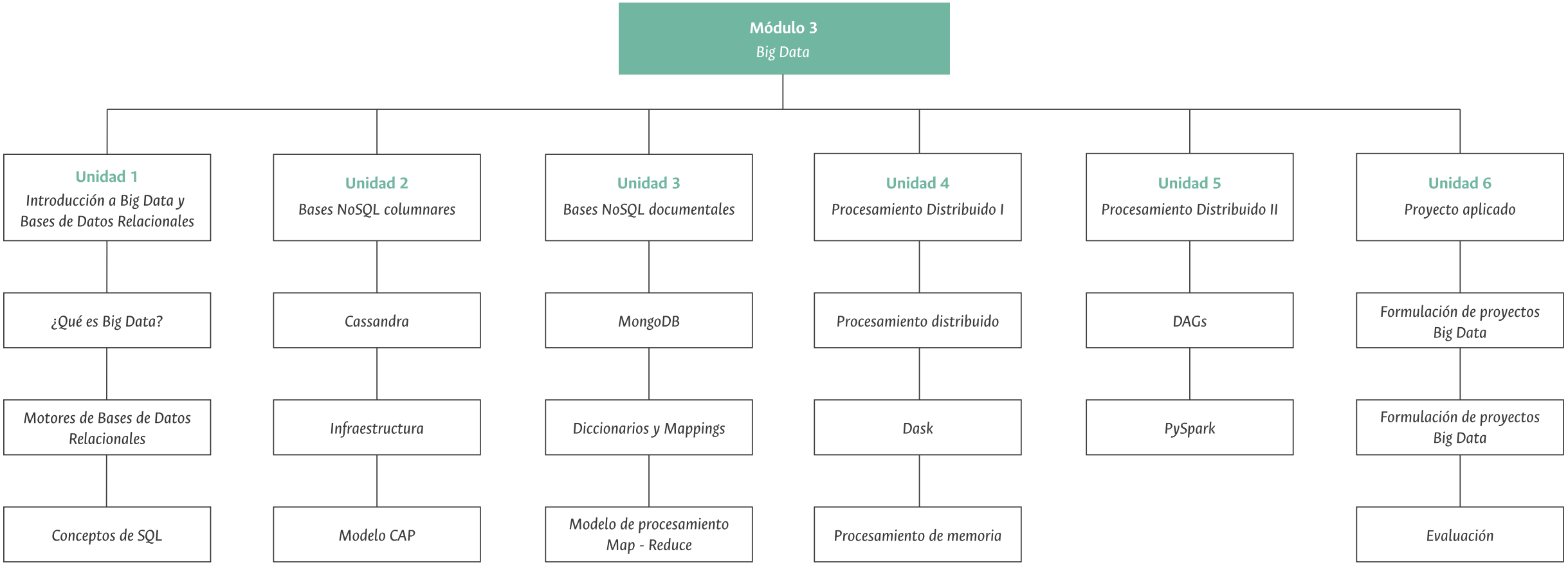
En este módulo aprenderá a desarrollar soluciones computacionales que involucren el almacenamiento, procesamiento y acceso a grandes volúmenes de datos, generados a velocidades vertiginosas y en diversos formatos. Al mismo tiempo obtendrá experiencia práctica en la selección de herramientas adecuadas de acuerdo a la naturaleza de distintos problemas.

Esperamos que disfrute la experiencia y que el conocimiento adquirido sea provechoso para progresar en su actividad profesional.

Meta

La meta de este módulo es proveer y capacitar al participante en conceptos y tecnologías para almacenar, procesar y acceder grandes volúmenes de datos, que son generados a grandes velocidades y en diversos formatos. Se busca que el participante conozca de manera teórica y práctica la arquitectura de manejadores de bases de datos *SQL y NoSQL*, el marco de trabajo distribuido *MapReduce*.

Mapa de contenidos del módulo



Objetivos de aprendizaje

Objetivos de aprendizaje

Unidad 1 - Introducción a *Big Data* y Bases relacionales

Al finalizar la unidad usted deberá ser capaz de:

Describir de manera precisa los conceptos generales de las distintas herramientas de almacenamiento de grandes cantidades de información y los distintos tipos de bases de datos.

Aplicar operaciones de creación, lectura, actualización y eliminación de datos con el lenguaje de consulta SQL estándar.

Utilizar motores de bases de datos SQLite y PostgreSQL desde Python.

Objetivos de aprendizaje

Unidad 2 – Bases NoSQL Columnares

Al finalizar la unidad usted deberá ser capaz de:

Describir las diferencias entre los distintos tipos de bases de datos *NoSQL* y su apropiada selección por medio del modelo CAP.

Entender el modelo de datos y la arquitectura en bases de datos columnares como *Cassandra*.

Utilizar el lenguaje de consulta CQL para manipular información desde la base de datos Cassandra y con el lenguaje de programación Python.

Objetivos de aprendizaje

Unidad 3 – Bases NoSQL Documentales

Al finalizar la unidad usted deberá ser capaz de:

Entender el modelo de datos y la arquitectura en bases de datos documentales como *MongoDB*.

Utilizar el lenguaje de consulta de *MongoDB* para manipular información desde el lenguaje de programación *Python*.

Entender conceptos de computación distribuida con operaciones de tipo *Map - Reduce* desde *MongoDB*.

Objetivos de aprendizaje

Unidad 4 – Procesamiento Distribuido I

Al finalizar la unidad usted deberá ser capaz de:

Entender conceptos de computación distribuida con operaciones optimizadas automáticamente por medio de grafos dirigidos acíclicos (DAGs).

Utilizar arreglos multidimensionales de forma distribuida y optimizada desde la librería *Dask* para el manejo de grandes cantidades de datos numéricos.

Utilizar *dataframes* distribuidos y optimizados desde la librería *Dask* para manejo de grandes cantidades datos tabulares.

Objetivos de aprendizaje

Unidad 5 – Procesamiento Distribuido II

Al finalizar la unidad usted deberá ser capaz de:

Entender conceptos de computación distribuida con operaciones optimizadas automáticamente por medio de grafos dirigidos acíclicos (DAGs).

Utilizar *dataframes* distribuidos y optimizados desde la librería *Spark* para el manejo de grandes cantidades de datos tabulares.

Objetivos de aprendizaje

Unidad 6 - Proyecto aplicado

Al finalizar la unidad usted deberá ser capaz de:

Ejecutar un proyecto en el que se requieran tecnologías *Big Data* de forma efectiva, mediante el uso de la metodología y las herramientas presentadas en el curso.

Palabras clave del módulo

Conceptos previos

- Programación básica en Python
- Uso de librerías científicas como:

NumPy

matplotlib

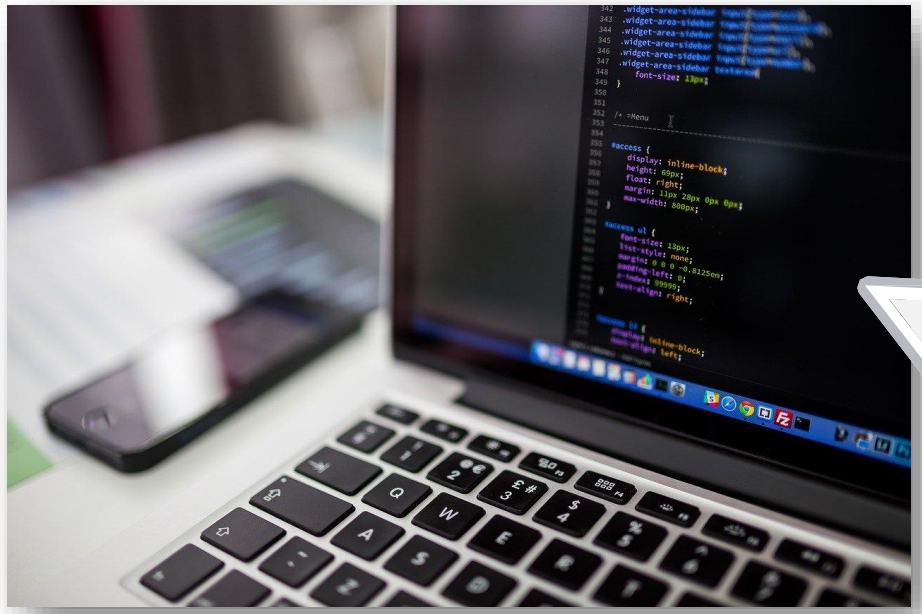
scikit learn

pandas

Habilidades y competencias a desarrollar

Habilidades y competencias a desarrollar

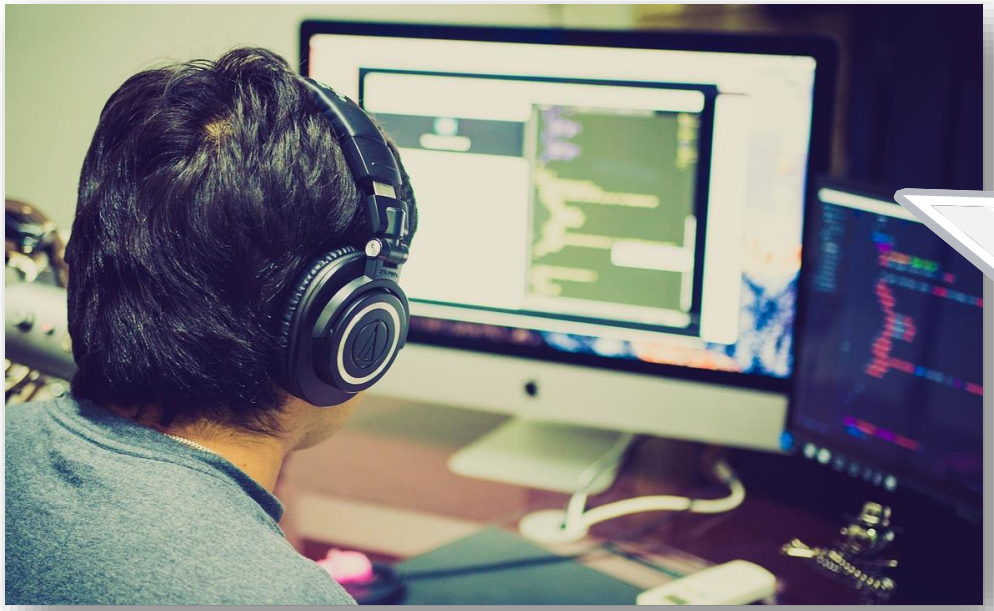
Saber



- Qué es *Big Data*.
- Principios fundamentales de las tecnologías de almacenamiento *Big Data*.
- Principios fundamentales de las tecnologías de procesamiento *Big Data*.
- Aplicaciones de *Big Data*.

Habilidades y competencias a desarrollar

Saber hacer



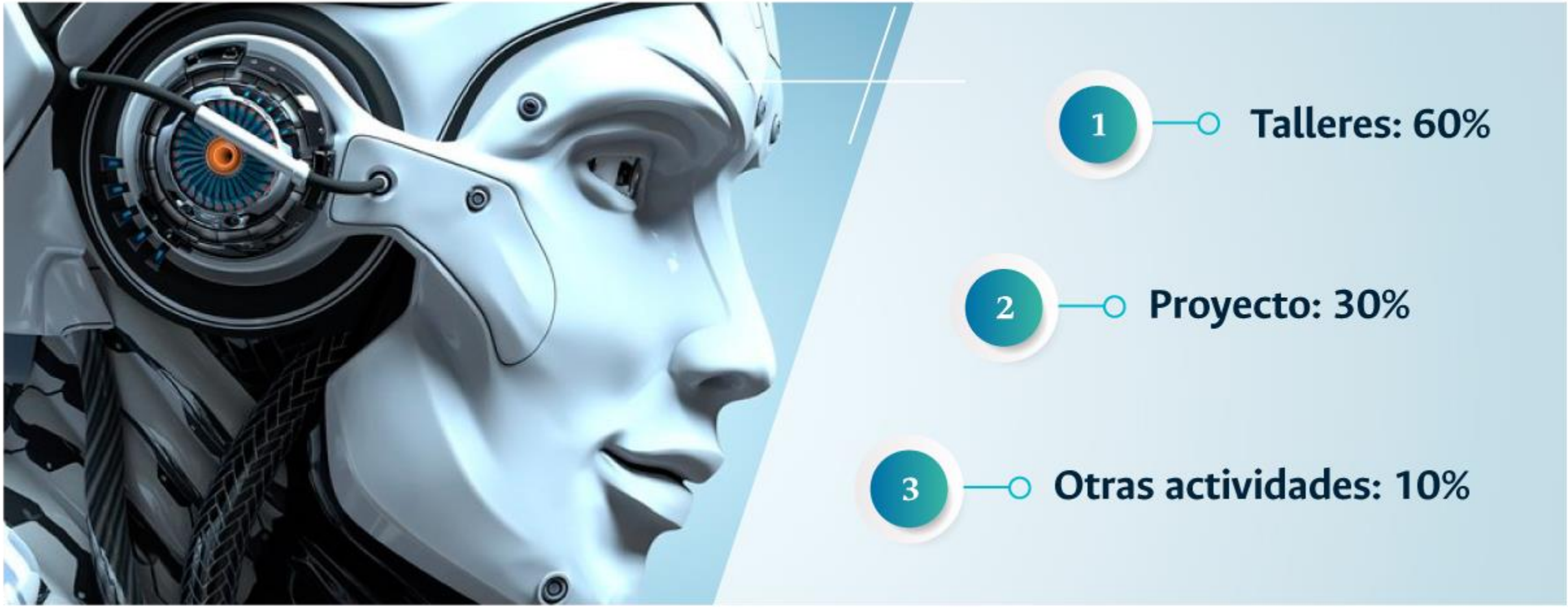
- Manejo de tecnologías de bases de datos *SQL* y *NoSQL*.
- Manejo de tecnologías para el procesamiento *Big Data*.
- Uso adecuado de tecnologías *Big Data* de acuerdo con el problema.

Habilidades y competencias a desarrollar

Saber ser

- Ganar confianza en el manejo de tecnologías para trabajar sobre grandes volúmenes de datos.
- Afianzar interés en el uso de tecnologías *Big Data*.

Evaluación formativa del módulo



¡Felicitaciones!

Ha concluido la introducción del tercer módulo del programa de formación MLDS: *Big Data*.

Le invitamos a continuar en la plataforma con la primera unidad del módulo.

Créditos

Facultad de
INGENIERÍA

Autores
Jorge Eliecer Camargo Mendoza, PhD

Asistente docente
Juan Sebastián Lara Ramírez
Edder Hernández Forero
Brian Chaparro Cetina
Rosa Alejandra Superlano Esquibel
Leonardo Avendaño Rocha
Alberto Nicolai Romero Martínez

Diseño instruccional
Claudia Patricia Rodríguez Sánchez

Diseño gráfico
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

Diagramadora PPT
Daniela Duque

Fecha
2022-II