|  |  |  |
| --- | --- | --- |
| **Nombre**: José Guillermo Robles Sánchez | | **Matrícula**: 2905526 |
| **Nombre del curso:**  Infraestructura de big data | **Nombre del profesor**:  Miguel de Jesús Martinez Felipe | |
| **Módulo**:  3 | **Actividad**:  Evidencia 2 | |
| **Fecha**: 2023 | | |
| **Bibliografía**: | | |

Introducción a la transformación y manipulación de datos

Apache Spark es un sistema de computación que tiene fundamentado a Hadoop MapReduce que permite fragmentar el trabajo y facilitarlo para poder procesar grandes cantidades de datos, ofrece velocidad computacional, escalabilidad y programabilidad.

Dentro de las ventajas de Spark es que facili uso de Python ya que cuenta con su propia librería.

La transformación de datos es un proceso común dentro del machine learning el proceso de convertir datos de un formato a otro del fuente al original, dentro de este proceso es necesario hacer otras acciones omo el refinamiento, combinaciones de datos recibiendo el nombre de ETL Extract, Load, Transform

Extract: se refiere a que los datos copiados vienen desde la fuente de origen

Load: estos datos se trasladan al almacen de destino

Transform: son los datos sin procesar o se modifican para su utilidad

Igual ecisten ciertos beneficios como:

-Alcanzar una mejor utilidad para los datos

-Administrar datos de forma eficiente

-Desarrollar consultas de forma mas rápida

-Aumentar la calidad de datos

Transformacion de datos en apacahe Hadoop

Spark tiene como fundamento procesar MapReduce pero manteniendo ciertas caraceristicas como su escalabilidad lineal y la resistencia a los fallos.

El DAG se refiere a un tipo de grafo que puede presentar series de datos relacionados los cuales son simbolizados por nodos en donde se representa un conjunto de datos dentro del grupo

El RDD facilita el desarrollo de operaciones que contienen grandes cantidades de volúmenes de los clústeres de forma rápida y eficaz, se realizan dos tipos de operaciones las cuales son: transformaciones y acciones

-tolerancia a fallos

-Computacion en memoria

-Estructuras de datos paralelas

-Inmutable

Existen distintos tipos de operaciones

Narrow transformation: se ubican en la partición que corresponden como filter(), sample(), map(), flatMap()

Wide transformation: estas combinan datos de diferentes particiones como groupByKey() y reduceByKey()

Dentro de las características de estas operaciones se destaca que

1. Tienen creación de valor mediante los datos
2. Tienen proyección de los datos
3. Tienen coherencia a los datos
4. Y se eliminan los datos que no sean necesarios