Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO VI: Frameworks de Procesamiento Masivo/Distribuido

Introducción:

En este trabajo se ejercitarán los conceptos fundamentales de Hadoop, con el modelo de programación MapReduce, y Spark, a través de PySpark. Para ello, trabajaremos con Python, emulando el modelo de programación de MapReduce con mappers y reducers y luego con PySpark para resolver problemas a partir de los RDD.

Hadoop MapReduce: El archivo *ventas.txt* posee las ventas de una Empresa, con los siguientes datos: Id_vendedor, Id_coordinador, Cantidad de productos vendidos y Cantidad de dinero.

Genere un esquema bajo el paradigma MapReduce para resolver las siguientes consignas:

a) Produzca un mapper y un reducer para responder a cuál es el bonus obtenido por cada vendedor siendo que cada vendedor obtiene el 3% del total del dinero vendido.
b) Produzca un mapper y un reducer para obtener la cantidad de productos vendidos por cada vendedor, agrupado por coordinador.

Apache Spark con PySpark: Resuelva el ejercicio anterior con PySpark.

Referencias sugeridas:

Hadoop: The Definitive Guide (4da edición) - Tom White. O'Reilly Media, Inc. 2015. ISBN: 978-1-491-90163-2.

Learning Spark - Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia. O'Reilly Media, Inc. 2015. ISBN 978-1-449-35862-4