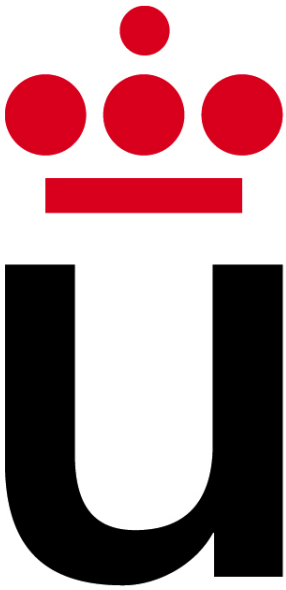
Foro de preguntas

Análisis de Big Data

Preguntas Tema 4

José Ignacio Escribano



Móstoles, 25 de marzo de 2016

Índice

1. Preguntas

1.1. Supongamos que cierto algoritmo de data mining se aplica sobre un gran conjunto de datos de 1PB de tamaño, almacenado en un sistema de ficheros distribuido en un cluster (por ejemplo, HDFS) y se va a analizar con un entorno que permite procesado de datos en memoria (por ejemplo, Spark). En cada iteración del algoritmo, se aplica una serie de pasos sobre un subconjunto de los datos, para luego, iterativamente, refinar el resultado en sucesivas pasadas muestreando de nuevo datos del conjunto original. ¿Sería necesario almacenar en este caso el total de 1PB de datos en memoria para agilizar los cálculos? ¿En qué principio nos podemos basar para argumentar nuestra respuesta?

1.2. Una alternativa al procesado de datos en memoria cuando no podemos almacenar toda la información en la RAM del sistema son las bases de datos de consultas aproximadas. BlinkDB, un proyecto de la UC Berkeley sigue esta estrategia. Busque información sobre BlinkDB y responda a estas preguntas: ¿Podemos acotar el error cometido en las consultas a este tipo de bases de datos? ¿Qué utilidad tienen respecto a las bases de datos tradicionales que siempre devuelven respuestas exactas a las consultas?¿Podemos acotar el tiempo de espera para recibir la respuesta a la consulta?