Master universitario in ANALISI DATI PER LA BUSINESS INTELLIGENCE
A.A. 2012-2013

Titolo della tesi: Big Data. Una nuova frontiera di data analysis. Vantaggi e limiti attraverso uno studio di caso.

Autore: Rocco Corriero

ABSTRACT

Esistono soluzioni di gestione dei dati ad alte prestazioni per l'analisi di dataset di grandi dimensioni. Questi stessi sistemi possono essere installati in ambienti data warehouse tradizionali per creare sistemi di gestione dei dati in grado di fronteggiare l'aumentare della quantità dei dati?
Sistemi di gestione basati su database paralleli sono in grado di fornire una soluzione ad alte prestazioni, ma sono costosi e complessi da implementare.
Lo scopo del progetto qui proposto è di confrontare la scalabilità dei sistemi di gestione di database relazionali open-source e sistemi di gestione dei dati distribuiti per grandi insiemi di dati in uso in ambienti di Big Data analytics.
Al fine di confrontare tali sistemi è stato messo a punto uno studio basato su un set di dati estratto da un server log e utilizzate tre soluzioni di gestione dei dati: MySQL, Apache Hive e Cloudera Impala.
L' esperimento ha coinvolto una sezione di un database di circa 18 milioni di righe.
Gli esperimenti sono stati eseguiti sottomettendo tre tipi di query: conteggio, lunghezza e aggregazione e registrando i tempi di esecuzione delle query stesse.
Nel nostro caso, con dati di tipo strutturato i risultati mostrano che MySQL ottimizzato per eseguire query su questo tipo di dati ha performance superiori rispetto Apache Hive. Discorso a parte va fatto con Cloudera Impala che si avvicina molto alle performance ottenute con MySQL. Una possibile spiegazione è costituita dal fatto che Impala è costruito su un'architettura proprietaria indipendente da Hive che aggira tutte le difficoltà incontrate da quest'ultimo costretto a tradurre le query in MapReduce. Cloudera Impala in questo senso rappresenta una valida alternativa nella gestione di grandi quantità di dati, in contesti esterni alla Big Data analytics.