Master Universitario in: “ANALISI DATI PER LA BUSINESS INTELLIGENCE”

A.A. 2018-2019

Titolo della tesi:

Analisi Cluster applicata a dati di stagionalità e di bilancio nel caso della grande distribuzione organizzata. Un prototipo di implementazione su piattaforma distribuita.

Autore: Matteo Malisan

Abstract

Lo stage finale del master ha fornito la possibilità di affrontare sul campo ed approfondire diverse questione tecniche relative ad ambienti di calcolo distribuiti (Hadoop e Spark innanzitutto) e di mettere in pratica almeno una parte delle tecniche di machine learning incontrate durante lo svolgimento del master.

Dal punto di vista dei software, ho potuto lavorare su di un cluster Spark e analizzarne la struttura e la configurazione (intervenendo anche su quest’ultima in modo da ottimizzare le prestazioni e sfruttare a pieno l’enviroment), e osservare le potenzialità di un file system distribuito come Hadoop e di una parte dei moltissimi software open source che la Apache Foundation sta raccogliendo e sostenendo (Hive, Zeppelin, Livy), oltre che dello stesso Spark. Nonostante la potenza limitata (si trattava di sole tre macchine), è stato possibile (confrontando le capacità di elaborazione con quelle di un classico database relazionale) testare in prima persona la scalabilità di questi sistemi. Inoltre, lavorando direttamente sulla shell di Linux, ho potuto approfondire notevolmente lo studio della struttura del cluster e anche installare del software aggiuntivo, che si è rivelato poi molto utile nel prosieguo.

Infine, la parte sperimentale ha portato a lavorare su una discreta mole di dati reali, serie storiche sulle quali si è potuto applicare tecniche di destagionalizzazione, smoothing e interpolazione polinomiale per poi, sfruttando Spark, effettuare diversi clustering dei dati, utilizzando K-Means ed anche tecniche come l’analisi delle componenti principali. I dati utilizzati sono stati i corrispettivi di vendita relativi a numerosi punti vendita di una nota catena internazionale ed i report trimestrali di valutazione degli stessi negozi, elaborato dall’ufficio controllo gestione dell’azienda nella quale si è svolto il tirocinio.

I risultati sono stati quelli desiderati e, soprattutto dal punto di vista dell’analisi delle serie storiche e del clustering basati sulla reportistica, anche interessanti.