Abstract

Scopo di questo lavoro, nato come progetto d'esame di basi di dati 2 tenuto dalla prof.ssa G. Tortora e G. Polese, è di esplorare le possibilità offerte dai big data, dalle caratteristiche che li contraddistinguono agli strumenti attualmente disponibili per acquisirli, memorizzarli ed estrarne informazione utile. Si è volutamente scelto uno stile pragmatico, per cui, relativamente alla tematica affrontata, saranno presentati numerosi esempi e contesti di utilizzo degli strumenti oggetti di studio. Pertanto, l'elaborato è rivolto alle aziende che siano intenzionate a valutare l'utilizzo di tali strumenti per accrescere il proprio business o a chiunque sia interessato ad intraprendere la carriera del data scientist "informatico, statistico e narratore che estrae pepite d'oro nascoste sotto montagne di dati", il mestiere che The Economist definisce "the sexiest job of the 21st century".

Conclusioni

Nel corso della trattazione abbiamo in prima istanza fornito le nozioni di base che contraddistinguono i big data: velocità, varietà, volume che spesso supera la reale capacità delle aziende di gestirli ed elaborarli con efficacia nei tempi utili. Una montagna di informazioni che costituiscono un problema se non utilizzati o usati poco o male, ma che possono trasformarsi in una formidabile opportunità quando vengono sfruttati nel modo corretto, ad esempio per sondare gli "umori" dei mercati e del commercio, e quindi del trend complessivo della società. Nel secondo capitolo abbiamo quindi appreso quali sono gli strumenti più idonei per "acquisire informazione", dalle API dei social network ai web scraper, cominciando ad introdurre Hadoop: un ecosistema di prodotti adottato da vendor del calibro di Facebook, Twitter durante l'intero life cycle dei big data. Nel terzo capitolo abbiamo inoltre appreso che, data la natura non strutturata dei big data quest'ultimi mal si prestano ad essere "trattati" con gli strumenti tradizionali quali gli RDBMS introducendo la tecnologia NoSQL e confrontando di volta in volta i vantaggi di taluna implementazione con le altre, compreso il modello relazionale.
Abbiamo evidenziato come strumenti tradizionali non siano obsoleti, sottolineando come i database relazionali rappresentino la scelta migliore per i sistemi OLTP, sia per quanto riguarda la "potenza nelle interrogazioni" che la loro capacità di garantire la consistenza nei dati. Gli strumenti presentati in questo survay quindi, completano il data warehouse, raramente lo sostituiscono. La maggior parte delle organizzazioni, infatti, ha progettato il proprio DW per dati strutturati e relazionali, il che rende difficile ricavare valore dalla BI con dati non strutturati e semistrutturati. Infine, nell'ultimo capitolo abbiamo presentato il framework MapReduce una potente architettura che nasconde i dettagli di parallelizzazione e

bilanciamento del carico, fault-tolerance, località dei dati, affinché, l'analisi dei big data, proibitiva fino a qualche anno fa, possa diventare una realtà per molte organizzazioni.