L'OPPORTUNITÀ BIG DATA: SFIDE IT E PROSPETTIVE PER LA STATISTICA UFFICIALE

di Domenico Donvito (donvito@istat.it)

La quantità di dati generata ogni giorno nel mondo ha un ordine di grandezza impressionante e deriva da un'incessante produzione da parte di molteplici sorgenti: rilevatori GPS, messaggi su social network, pagine internet, file di log, dati di esperimenti scientifici e altro ancora.
Il termine "Big Data" è utilizzato per identificare insiemi di dati strutturati o destrutturati che, per le loro dimensioni non sono gestibili utilizzando i comuni sistemi software (come i database relazionali).
La disponibilità di grandi quantità di dati, al di là dei vincoli tecnici, apre possibilità per realizzare nuove analisi ed estrarre nuova informazione. Esempi si riscontrano in vari domini applicativi, dalla sanità alle scienze sociali, dove l'analisi di grandi data set permette di svelare informazione altrimenti nascosta, ad esempio sotto forma di cluster e pattern.

BIG DATA PER LA STATISTICA UFFICIALE

L'utilizzo dei Big Data nella statistica ufficiale presenta varie criticità che sono legate da un lato all'intrinseca complessità di questo paradigma di gestione dei dati e dall'altro alla natura "ufficiale" dei prodotti forniti dagli INS.
Dimensioni. La memorizzazione di Big Data richiede architetture avanzate supportate da tecnologie complesse come file system distribuiti (Hadoop), database machines (Oracle Exadata), database NoSQL (Cassandra) e verticali. Nonostante queste soluzioni siano ampiamente disponibili sul mercato, il loro utilizzo negli INS è attualmente limitato.
Qualità. I Big Data pongono questioni

di qualità a vari livelli: qualità del dato (correttezza, aggiornamento, completezza), affidabilità o reputazione della sorgente e metadati. I Big Data sono anche "noisy", perciò richiedono una fase preliminare di filtraggio e pulizia. Quindi una problematica importante riguarda il bilanciamento tra la complessità del processo di pulizia rispetto al valore dell'informazione ottenuta.
Aspetti temporali. I Big Data sono altamente volatili, nel senso che l'informazione che rappresentano diventa rapidamente obsoleta, e variano in maniera continua. Questo rende difficile il loro utilizzo tempestivo e corretto nelle analisi.
Accessibilità. Il ruolo principale che dovrebbero avere gli INS rispetto ai Big Data è quello di consumatori piuttosto che di produttori. I produttori principali di Big Data (ad esempio Google, Facebook, ecc.) vedono il loro principale vantaggio competitivo proprio nel poter disporre in maniera esclusiva dei loro dati. Anche una volta risolto il problema dell'accesso, la privacy può essere un altro fattore che limita l'accesso ai dati.
Nonostante le criticità sopra esposte, l'abbondanza di dati nella società odierna dovrebbe stimolare una profonda ed ampia riflessione sui prodotti offerti dalla statistica ufficiale considerando anche il rischio che gli istituti nazionali di statistica possano essere superati da fornitori non ufficiali, più pronti a trarre vantaggio dalle nuove opportunità.

INIZIATIVE IN ISTAT

L'Istat ha cominciato un processo di adozione delle tecnologie Big Data, di cui sono esempi la migrazione del patrimonio di dati verso una database machine Oracle Exadata e l'utilizzo della tecnologia PowerPivot per le attività di Business Intelligence (BI) per il Censimento della Popolazione. Inoltre, è stato avviato un accordo di collaborazione con il Dipartimento di informatica dell'Università di Pisa, con potenziali aperture anche a livello Europeo, per lo studio di nuove tecniche ed è in fase di costituzione una Commissione scientifica dedicata allo studio del tema costituita sia da membri Istat che da esperti del mondo accademico e industriale.
Per concludere, le potenziali opportunità fornite dai Big Data sono tali da spingere a fronteggiare le sfide elencate.
I metodi e le tecnologie Big Data potrebbero infatti portare non solo ad una più efficiente produzione statistica, ma addirittura a cambiamenti sostanziali nel modo di produrre le statistiche, sfruttando la varietà, la ridondanza e la disponibilità "near-time" dei Big Data.