Introduzione ai Big Data

Francesco Pugliese, PhD

neural1977@gmail.com

Disponibilità di una quantità massiccia di dati

- ✓ I dati digitali sono oggigiorno collezionati ad una scala **senza precedenti** ed in molti formati **in una varietà di domini** (e-commerce, social network, reti di sensori, astronomia, genomica, registri medici, ecc.)

- ✓ Questo è stato reso possibile da un'incredibile crescita negli ultimi anni della **capacità degli strumenti di data storage** e della potenza **computazionale dei dispositivi elettronici**, come anche con l'avvento della computazione mobile e pervasiva, cloud computing e cloud storage

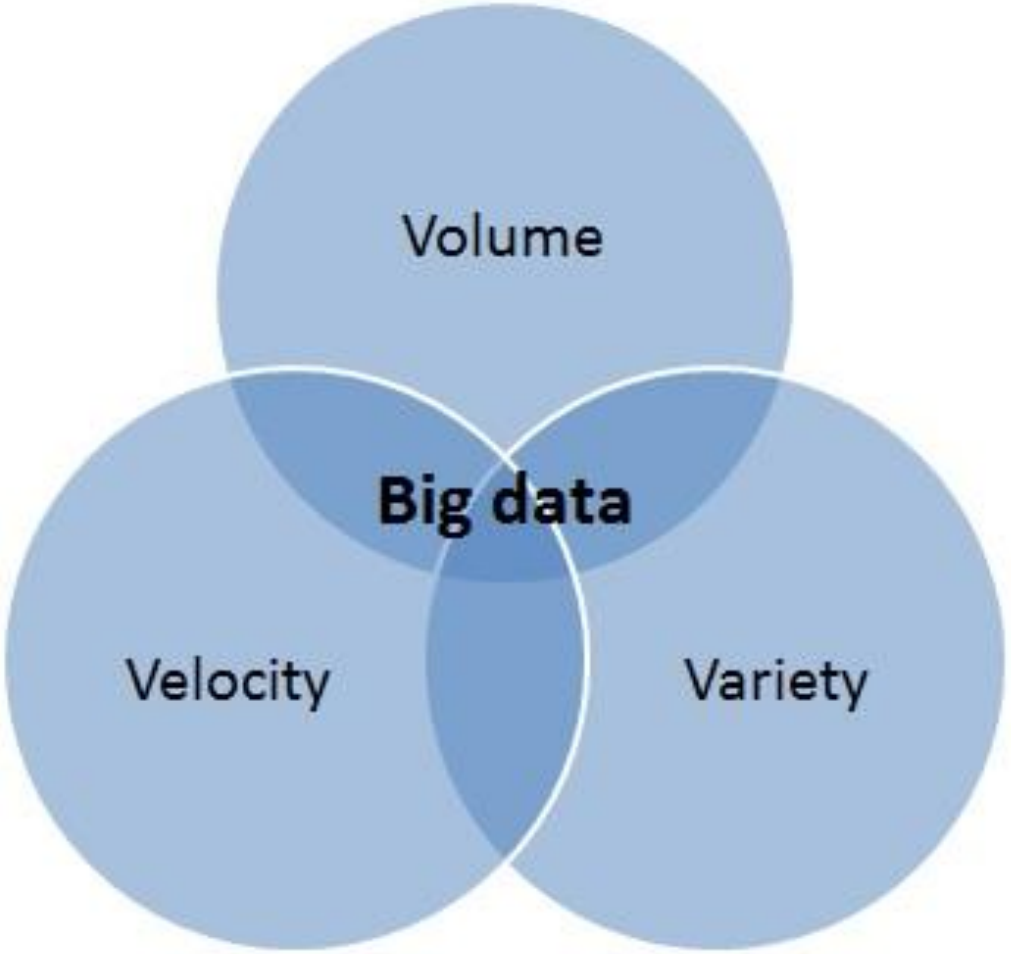
Sfruttabilità di Dati Massivi

- ✓ Come **trasformare i dati disponibili in informazione,** e come creare il business delle organizzazioni per trarre vantaggi da tale informazione è un problema di lunga data nell'IT, e in particolare in sistemi di analisi e gestione dell'informazione.
- ✓ Queste questioni sono diventate sempre più sfidanti e complesse nell'era dei **Big Data.**
- ✓ Allo stesso tempo, affrontare la sfida può essere anche più di valore che nel passato, dal momento che una quantità massiccia di dati cioè ora disponibile potrebbe permettere risultati analiti mai raggiunti prima.

Le tre V

- ✓ Per caratterizzare i Big Data, sono usate le 3 V, che sono le V di:

- ✓ - Volume
- ✓ - Velocity
- ✓ - Variety



Volume

- Le applicaziojne dei Big Data sono caratterizzate certamente da grandi quantità di dati, dove big significa **estremamente grandi**, per esempio terabyte o petabyte o più.
- Ci sono vari contesti in cui queste dimensioni possono essere facilmente raggiunte: chatters dei social networks, log dei web server, sensori a basso traffico, immagini satellitari, flussi di audio di broadcast, transazioni bancarie, dati di mercati finanziari, dati biologici, ecc.
- Alcuni esempi più concreti:
- Malgrado le statistiche di Youtube siano disponibili, la capacità di storage totale di Youtube non è conosciuta, ma realisticamente dovrebbe non essere meno di un 1 EB.
- NSA data center 2000 PB, Facebook 300PB

Volume

- ✓ Il **solo** volume dei dati è abbastanza per far fallire molti degli approcci presenti da decenni nel data management
- ✓ I database centralizzati tradizionali non possono gestire molti dei volumi dei dati, forzando l'uso dei cluster
- ✓ I dati devono essere necessariamente distribuiti, e il **numero delle sorgenti che forniscono informazione possono essere enormi,** molti più alti del numero considerato nella tradizionale integrazione dei dati e sistemi di virtualizzazione

Velocity

- ✓ La Velocity dei Dati, ossia il tasso con cui i dati sono collezionati e resi dispobibili per un'organizzazione, ha seguito un percorso simile a quello del Volume.
- ✓ Molte delle sorgenti dei dati accedute dalle organizzazioni per il loro business sono estremamente dinamiche
- ✓ I dispositivi mobili incrementano il tasso del flusso entrante di dati: dati "ovunque", collezionati e consumati continuamente

Velocity

- ✓ **Elaborare l'informazione non appena essa è disponibile,** quindi velocizzare il "feedback loop", può fornire vantaggi competitivi
- ✓ Alcuni esempi di **Elaborazione dei Dati Veloce:**
- ✓ Customer Experience/Retail: retail online che sono abili a suggerire prodotti addizionali a un cliente a ogni nuova informazione inserita durante un acquisto on-line
- ✓ Industria dei Servizi Finanziari: Trading algoritmico usando una tecnologia di elaborazione degli eventi ma anche data integration real time e analisi
- ✓ Telecomunicazioni: comprendere l'allocazione di risorse di rete basate su traffico e requisiti di applicazione, schemi di uso della rete

Velocity

- ✓ Lo **Stream Processing** è un nuovo paradigma computazionale molto sfidante, dove l'informazione non viene immagazzinata per il batch processing successivo, ma viene consumato al volo
- ✓ Questo è particolarmente utile quando **i dati sono troppo veloci per immagazzinarli interamente** (per esempio a causa del fatto che essi hanno bisogno di determinata elaborazione per essere immagazzinati propriamente), come nelle applicazioni scientifiche, o quando l'applicazione richiede una risposta immediata.

Variety

- I dati sono estremamente eterogenei: cioè nel formato in cui sono rappresentati, ma anche e nel modo in cui essi rappresentano l'informazione, entrambi ad un livello intenzionale ed estensionale.
- Ad esempio, il testo dai social network, dati di sensori, log dalle applicazioni web, database, documenti XML, dati RDF, ecc.
- Il formato dei dati varia quindi dallo strutturato (cioè database relazionali) a semistrutturati (cioè, XML, documenti), a non strutturati (cioè, documenti di testo)

Una quarta V: Veracity

- ✓ I dati sono ampiamente differenti in qualità
- ✓ I dati sono tradizionalmente pensati per provenire da database ben organizzati con schemi controllati
- ✓ Invecem nei Big Data c'è spesso poco o nessuno schema di controllo della loro struttura
- ✓ Il risultato è che ci sonno seri problemi con la qualità dei dati.
- ✓ La letteratura spesso menziona solo le **tre V** e non include la veracity. Tuttavia, alcuni autori tendono a includere veracity come una caratteristica core dei Big Data (alternativamente, la veracity è considerata un aspetto della variety).

Big Data: V3+Value

- ✓ **I Big Data possono generare grandi vantaggi competitivi!**