< (BD∪BI) ∧ ↘€>

Micaela Caserza Magro, Paolo Pinceti
DITEN – Università degli Studi di Genova
Via Opera Pia 11a – 16145 Genova
Tel. 010/3532205 - Email: paolo.pinceti@unige.it

Introduzione

Cosa significa il titolo di questo lavoro? In che lingua è scritto? E' vero quello che dice?

Una traduzione in italiano corrente del titolo potrebbe suonare come "Big Data insieme a Business Intelligence implicano minori costi". Il linguaggio usato è un gramelot tra logica matematica, mania degli acronimi che stiamo importando dagli americani, XML, fantasia pura. La domanda fondamentale (e seria) è la terza: è vero che le tecnologie "Big Data" e la "Business Intelligence" possono portare benefici all'automazione industriale?

Ma prima di rispondere a questa fondamentale domanda, cerchiamo di chiarire cosa significhi realmente big data, al di là dei discorsi "un po' da bar" che spesso si fanno sul tema. Si parla di Big data quando i dati da gestire –nel senso di acquisire, immagazzinare, elaborare, ricercare,…- superano la capacità dei normali database. In questo senso la soglia dei Big Data si muove giorno per giorno, alzandosi insieme all'aumento di capacità dei database: a seconda dell'interlocutore si può passare dalle centinaia di gigabytes (10^{11}) alle migliaia di terabytes (10^{15}). Chiariamo quindi sin da subito che nessun sistema industriale può generare moli di dati di queste dimensioni, a meno che il progettista del sistema di automazione non abbia commesso qualche madornale errore di programmazione.

Allo stesso modo, la Business Intelligence è un insieme di tecniche statistico-matematiche utili a trasformare insiemi di dati grezzi in informazioni utili alla gestione dell'azienda. Le tecniche di BI nascono, come dice il nome, nel settore economico-finanziario, con applicazioni oggi prevalentemente di marketing, ed hanno nomi evocativi come: benchmarking, data mining, business performance management, ecc. ecc.

Appare evidente che BD e BI non sono *sic et simpliciter* strumenti idonei ad applicazioni nell'ambito dell'automazione dei processi industriali e delle macchine, ma è altrettanto evidente che possono essere mutuate da questi ambiti alcune delle metodologie che possono essere utilizzate proficuamente in quest'ambito. Questo lavoro ha proprio questo scopo: identificare che tecniche e approcci possono essere impiegate per analizzare la massa di dati (non Big) che i sistemi di automazione integrati rendono oggi disponibili.

Tecniche Big Data e Business Intelligence

I cosiddetti Big Data sono associati alle tre “V”:

- Volume: quantitativi di dati non gestibili con i database tradizionali,

- Velocità: dati che fluiscono e devono essere processati in tempo reale,

- Varietà: dati eterogenei tra loro sia come tipologia (video, testi, audio,…) sia come sorgente di provenienza.

C’è chi aggiunge una quarta “V” per Visualizzazione, cioè le modalità di presentazione dei risultati dell’analisi all’utente. Altri aggiungono un’altra “V” ancora: Veridicità o Validazione. Se le 3V (o quattro o cinque che siano) sono il marcatore dei Big Data, è chiaro che non si deve parlare di Big Data nel mondo dell’automazione industriale. Infatti Big Data nasce nei settori legati a Internet, in particolare nel mondo consumer, ed è bene che li resti. Proprio l’esigenza di operare su enormi quantità di dati ha fatto si che BD sviluppasse piattaforme hardware e pacchetti middleware propri e specifici. La Fig. 1 mostra un esempio (di IBM) di architettura idonea alla gestione di BD, basata su processori massivamente paralleli (tecnologia blade), e grandi banchi di dischi. Gli SMP Host (Simmetric Multi Processing) rappresentano l’interfaccia verso l’esterno ed il cervello di gestione dei processori blade.

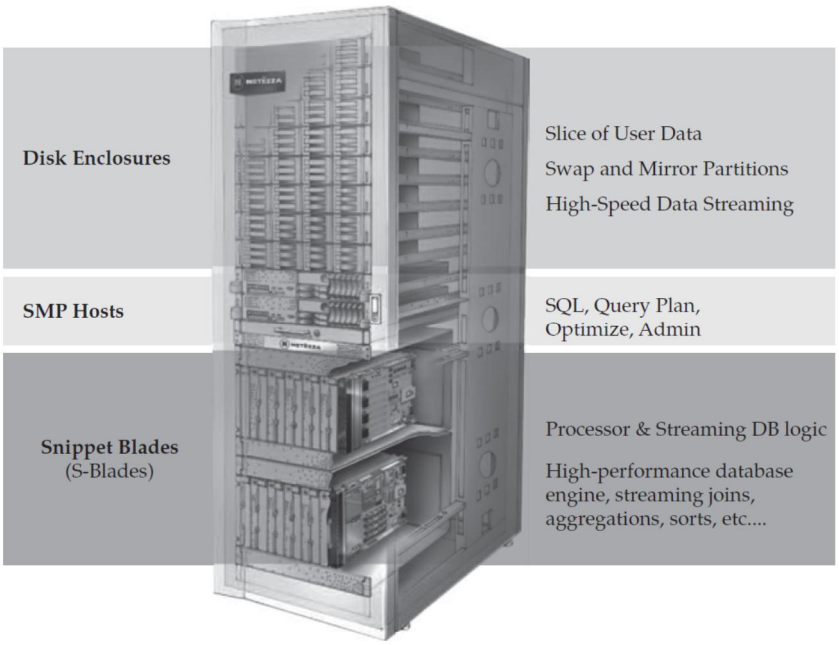


FIG.1 TIPICA ARCHITETTURA HARDWARE PER LA GESTIONE DI BIG DATA

L’automazione industriale costituisce un ambiente informatico volutamente e fortemente omogeneo, caratterizzato da flussi di dati perfettamente gestibili dai database tradizionali. Le analisi dei dati del processo sono inoltre legate a funzionalità dove i vincoli di real time sono molto limitati o assenti: la diagnostica, la manutenzione, la gestione degli asset,…

La Business Intelligence può essere definita come l'insieme delle metodologie, modelli, tecnologie e applicazioni rivolte alla raccolta sistematica del patrimonio di informazioni generate ed acquisite da un'azienda, alla loro aggregazione, analisi, presentazione ed utilizzo. Esistono numerose piattaforme informatiche sulle quali implementare sistemi di BI, alcune proprietarie, altre open source (ad esempio la BI Suite di Jaspersoft). Un'applicazione di BI ha tipicamente una struttura del tipo di quella riportata in Fig.2.

FIG.2 TIPICA STRUTTURA DI UN'APPLICAZIONE DI BI

I dati vengono raccolti da sorgenti eterogenee ed immagazzinati in un database relazione tradizionale; è la fase di Extraction, Transformation, and Loading (ETL) dei dati. Per velocizzare la fase di analisi vengono spesso utilizzate tecniche di On-Line Analytical Processing (OLAP) che prevedono la ridefinizione del database in una struttura multidimensionale più rapidamente accessibile del database relazionale di partenza. In breve, partendo dai dati non aggregati del database vengono predefinite le loro possibili aggregazioni ed i dati sono ri-immagazzinati secondo queste. Ad esempio, il DB ha dati su impianti, apparati, e strumenti, ed ognuna di queste categorie contiene parametri di funzionamento. Con l'OLAP i dati di funzionamento sono raggruppati sulla base di qualsiasi combinazione (impianti, apparati, strumenti), o meglio sulla base di qualsiasi combinazione ritenuta interessante. La struttura più comune per l'organizzazione dei dati è quella a stella (cfr. Fig.3). La "tabella dei fatti" rappresenta il database relazionale di partenza che contiene, ad esempio, le misure effettuate. Questa è associata a N tabelle dimensionali ciascuna delle quali mantiene le relazioni gerarchiche originarie (ad esempio anno/mese/giorno). Perché il sistema funzioni, il database deve essere riportato su almeno due dimensioni, ma non su troppe. E' necessario scegliere le dimensioni con attenzione, sulla base delle future analisi che saranno condotte sui dati. Dalla struttura a stella si costruisce un cubo multidimensionale che contiene i dati ("fatti" nel linguaggio del mondo OLAP) già organizzati secondo le dimensioni scelte. E' sul cubo che vengono fatte le query per l'estrazione dei dati.

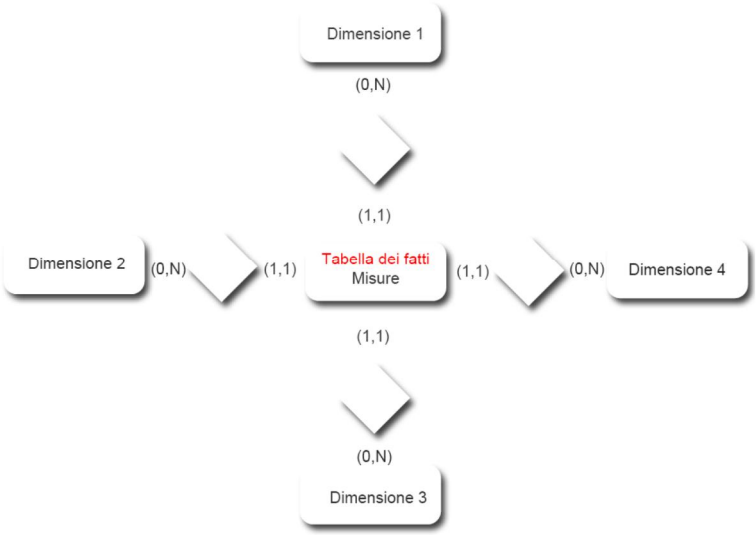


FIG.3 STRUTTURA A STELLA DI UN DATABASE OLAP

I dati dell'Automazione Industriale

Nei sistemi di automazione integrati, che utilizzano cioè Intelligent Field Device (IFD) o Intelligent Electrical Device (IED) e connessioni su fieldbus, abbiamo a disposizione per ciascun dispositivo in campo e per ciascun sottosistema una serie di dati che vanno ben oltre i soli parametri di processo. Ed è proprio questa disponibilità di dati che permette di applicare tecniche di analisi matematico-statistiche in grado di ottenere informazioni utili alla diagnostica ed al monitoraggio del processo.

Il punto di partenza è rappresentato dalla struttura dati che rende disponibile ciascun oggetto o ciascun sottosistema. Ogni elemento presente sulla rete di comunicazione viene virtualizzato e visto come un contenitore di dati, organizzati secondo attributi particolari o funzionalità. Un buon esempio di struttura dati è offerto dalla tecnologia GOMSFE del protocollo IEC 61850[1], specifico per il mondo elettrico, per il quale ogni apparati fisico è un server che rende disponibile ai client che vi si connettono i dati raggruppati in "nodi logici" specifici per ciascuna funzione (misura, protezione, controllo, ecc.). Ogni nodo logico contiene a sua volta delle classi di "dati" che al loro interno hanno i valori atomici di informazione (attributi). Questa struttura ad albero è del tutto identica alla struttura a cartelle cui siamo abituati nella gestione dei file immagazzinati in un disco del PC (vd. Fig.4).

Un sistema composto da molti Intelligent Device è quindi costituito da analoghe strutture dati che si ripetono. IED diversi conterranno dati diversi, ma la struttura è comunque analoga. Nel mondo del processo la struttura dati è diversa da quella prevista dalla IEC 61850, ma anche qui si trovano "profili" tipici dei diversi apparati che contengono dati all'interno di classi.

[1] Generic Object Model for Substation and Feeder Equipment

I contenitori di dati possono essere sia singoli apparati, come sin qui visto, sia sottosistemi o macchine. Quel che conta è costruire una base dati con una numerosità sufficiente ad identificare trend e comportamenti.

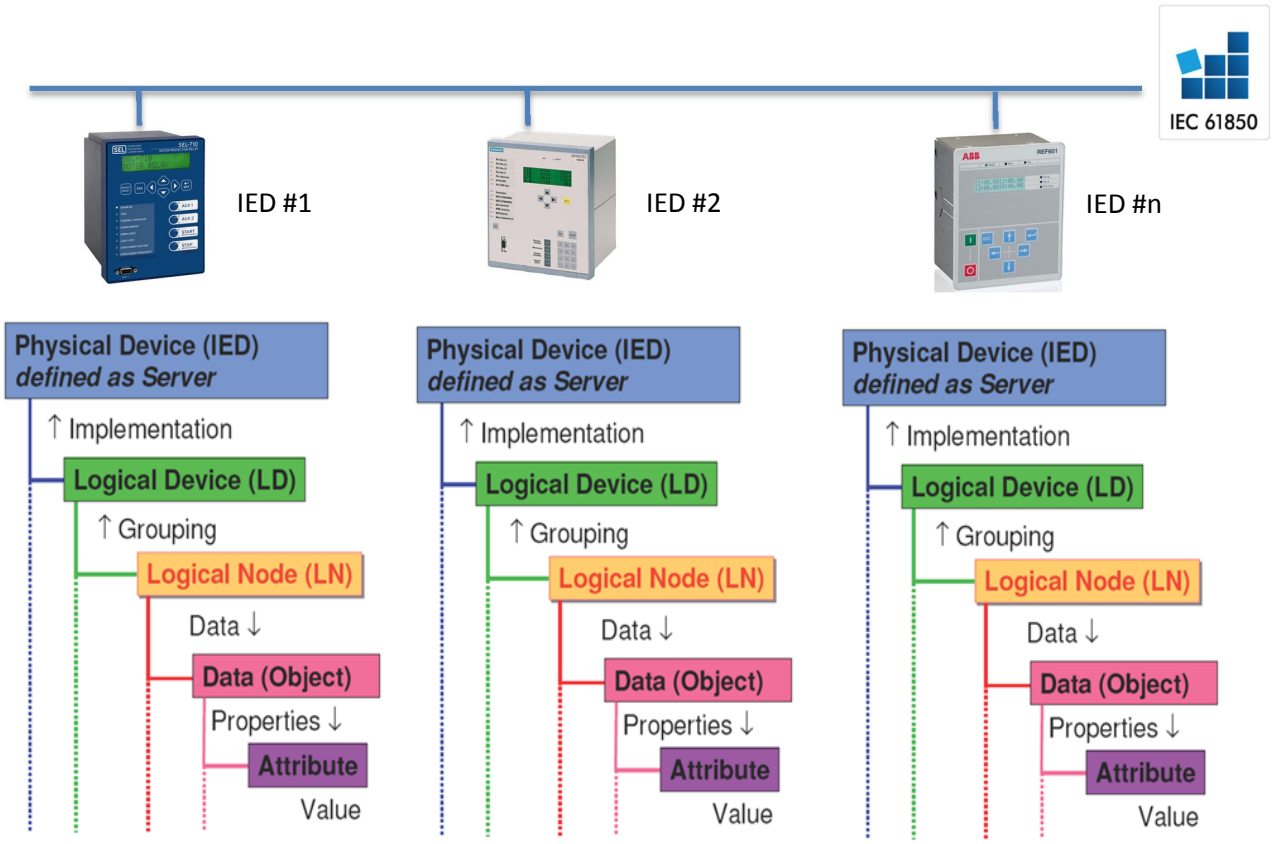


FIG.4 STRUTTURA DATI SECONDO IEC61850

Le analisi matematico-statistiche dei dati

Partendo da una base dati distribuita o concentrata ma caratterizzata comunque da una struttura a matrice dove le colonne sono gli apparati e le righe le classi di dati che contengono (vedi Fig.5), è possibile identificare tre tipologie di analisi:

1) Orizzontale: analisi di proprietà uguali di apparati diversi,
2) Verticale: analisi dell'andamento di una o più proprietà di un singolo apparato
3) Mista: analisi incrociata di proprietà diverse dello stesso apparato o di apparati diversi.

Macchina/Impianto	#1	#2	#3	…	#m
Parametro #1	p_{11}	p_{12}	p_{13}	…	p_{1m}
Parametro #2	p_{21}	p_{22}	p_{23}	…	p_{2m}
Parametro #3	p_{31}	p_{32}	p_{33}	…	p_{3m}
…..	…	…	…	…	…
Parametro #n	p_{n1}	p_{n2}	p_{n3}		p_{nm}

FIG.5 ESEMPIO DI BASE DATI PER "M" MACCHINE MONITORATE SU "N" PARAMETRI

La scelta del tipo di analisi deve essere compiuta dal progettista sulla base della specificità del suo impianto/sistema. In generale è vero che:

a) aziende che costruiscono macchine o impianti ripetitivi potranno trarre vantaggio da un'analisi orizzontale che compari la stessa proprietà tra macchine/impianti diversi. Per questo può essere utile definire indici di prestazione (ad es. il consumo specifico per unità di prodotto, la producibilità della macchina, il numero di fermate, ecc.) che diano un'immediata visione dello stato di funzionamento della macchina/impianto. La comparazione orizzontale identifica immediatamente comportamenti anomali (vd. Fig.6) di una macchina/impianto rispetto ai valori tipici, ma può offrire anche spunto al progettista per analisi di tipo misto per cercare di comprendere se un'anomalia di un parametro potrebbe essere correlata ad altre caratteristiche. La semplice analisi statistica di varianza aiuta questo tipo di analisi: la fig.6 mostra l valore di un parametro su dieci macchine o impianti diversi. Evidenziando la banda di varianza su ciascuna valore è immediato identificare possibili anomalie da investigare (macchina 4 e 9) o da monitorare (macchina 6);

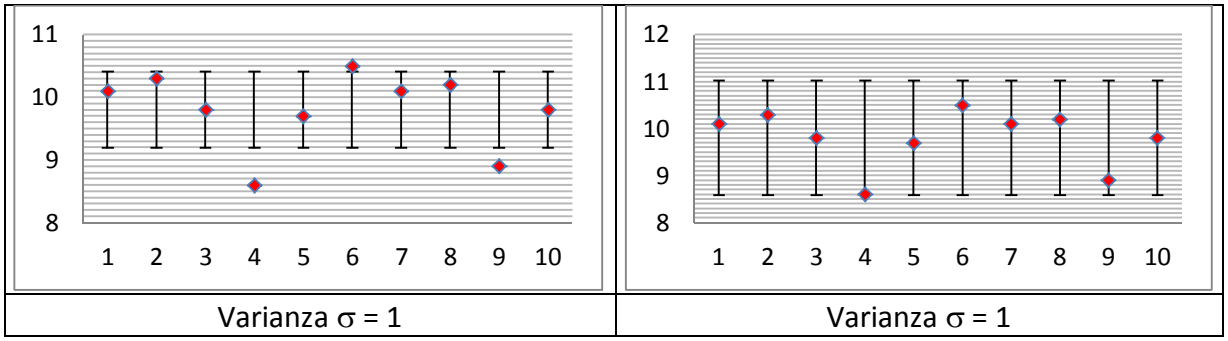
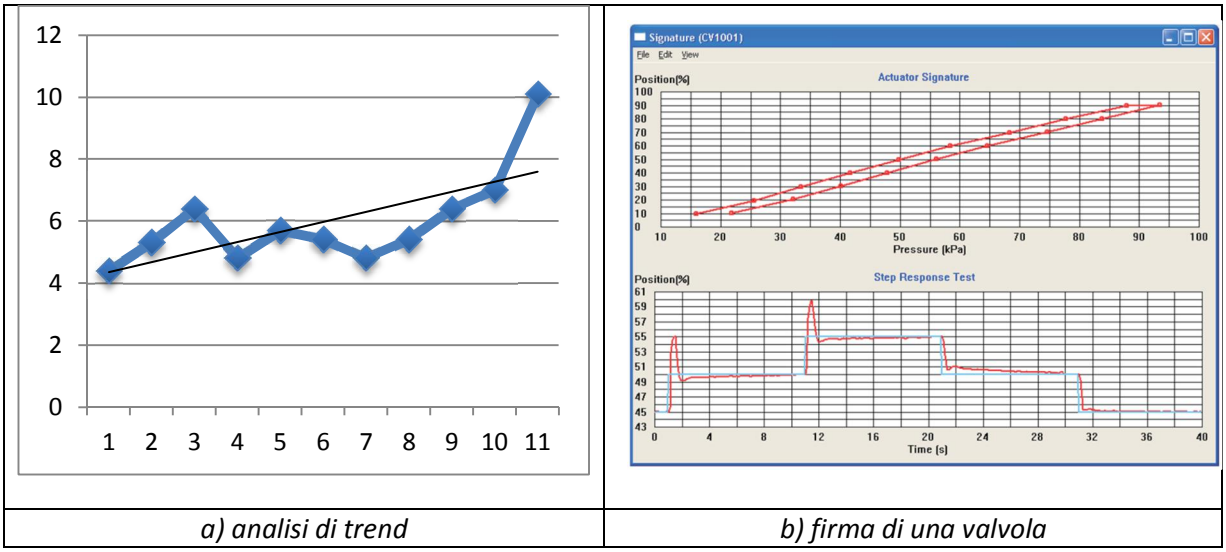


FIG.6 ESEMPIO DI ANALISI DI VARIANZA DI UN PARAMETRO SU DIECI MACCHINE/IMPIANTI

b) impianti o macchine singoli o comunque non riconducibili ad un progetto unitario non traggono benefici da un'analisi orizzontale (per troppo bassa numerosità dei campioni), ma possono utilizzare l'analisi di tipo verticale. Questa è orientata allo studio dei dati della singola macchina/impianto attraverso tecniche di:

- analisi del trend temporale di un parametro o di un indice di prestazione (Fig.7.a). Uno scostamento eccessivo dal trend può indicare ;
- analisi della firma di un apparato (per "firma" si intende l'andamento di un determinato parametro al ripetersi di un'operazione nota, ad esempio l'andamento della coppia di un posizionatore di una valvola on/off al comando di apertura/chiusura in Fig.7.b). Scostamento della firma dalla sua forma solita sono indice di insorgenza di una modifica funzionale che deve essere investigata;
- analisi delle anomalie, anche in questo caso operando su un orizzonte temporale il più lungo possibile, per evidenziare aumenti repentini o lenti di eventi o parametri riconducibili ad anomalie.



a) analisi di trend | *b) firma di una valvola*

FIG.7 ESEMPIO DI ANALISI VERTICALE

c) l'analisi mista del database è applicabile a tutte le macchine/sistemi ed è più che altro rivolta al progettista della macchina/impianto o a chi ne utilizza o gestisce più di uno. L'analisi mista può essere condotta su parametri diversi della stessa macchina/impianto (misto-verticale) o anche di macchine diverse (misto-incrociato). Scopo fondamentale dell'analisi mista è quello di evidenziare correlazioni tra parametri o eventi diversi, identificando comportamenti non immediatamente o preventivamente noti. Le tecniche statistiche adottate in prevalenza per queste analisi sono:

.regressione: si tratta di determinare una funzione matematica che esprima la relazione tra le variabili scelte, dopo un'analisi di ragionevolezza del loro presunto legame[2]. La funzione di regressione è il più delle volte ottenuta utilizzando il metodo dei minimi quadrati;

. correlazione: la correlazione tra due (o più) variabili si misura mediante indici che esprimono l'intensità del loro legame, solitamente normalizzati nel range da -1

[2] l'analisi di regressione tra variabili eterogenee può portare a risultati assurdi. Si ricordi il legame trovato, in epoca vittoriana, tra l'incidenza della tisi e l'uso del cappello a tuba

(correlazione perfetta negativa) a +1 (correlazione perfetta positiva). L'indice più comunemente utilizzato è il "coefficiente di correlazione di Pearson", espresso con la lettera "r" o "ρ". Come mostra la Fig.8, nel caso di due variabili, già la semplice rappresentazione nel piano cartesiano può mostrare l'esistenza di correlazioni. Nella ricerca di correlazioni tra variabili diverse occorre prestare molta attenzione al fenomeno delle "correlazioni indirette", dove cioè due variabili sono correlate attraverso una terza non esplicita[3].

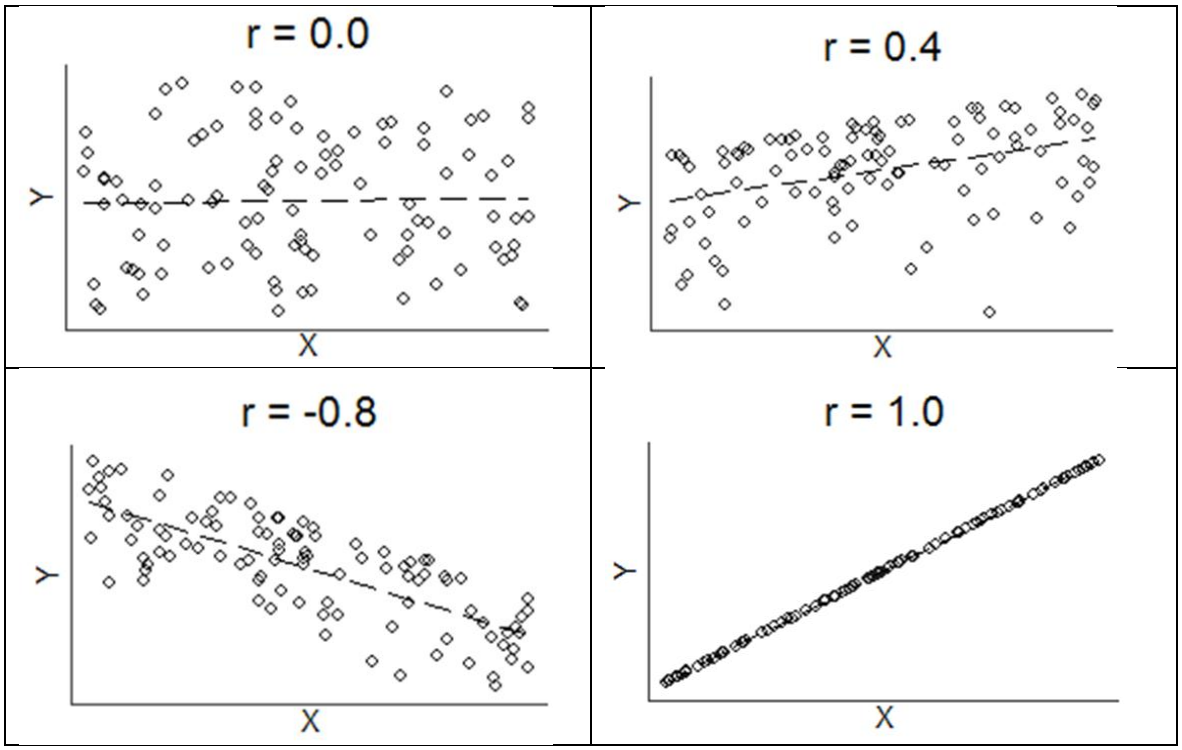


FIG.8 ESEMPIO DI CORRELAZIONE D DUE VARIABILI (X, Y)

[3] tipico esempio: il "numero delle vittime in un incendio" (variabile X1) è correlato direttamente al "numero di pompieri impiegati per spegnerlo" (variabile X2), da cui sembrerebbe logico mandare pochi pompieri a spegnere gli incendi. In realtà X1 e X2 dipendono entrambe dalla terza variabile "dimensione dell'incendio".

Conclusioni

Pur essendo certamente vero che i sistemi di automazione integrati, basati cioè su reti di Intelligent Device, possono rendere disponibili grandi quantità di dati, non sembra proprio necessario scomodare i Big Data. I normali database relazionali, magari accoppiati a tecniche di gestione dei dati mutuate dalla Business Intelligence, possono sicuramente essere utilizzati per implementare analisi di tipo statistico/funzionale utili ad identificare anomalie di funzionamento, difettosità, incipienti guasti, in macchine o impianti. A seconda della numerosità possono essere utilizzate tecniche di analisi di tipo verticale (sui dati storicizzati di una singola macchina/impianto), orizzontali (tra macchine/impianti diverse), o miste. Queste ultime analisi possono portare all'identificazione di correlazioni funzionali non immediatamente evidenti o note a priori.

Bibliografia

[1]

[2]

[3]