



Data Engineering Essentials: Governare il Dato per Creare Valore

Il Dato: cosa ci portiamo a casa dopo oggi

Ing. Francesco Micillo
*Presidente Commissione Informatica
Ordine degli Ingegneri della Provincia di Roma*

Roma, 01 luglio 2025



Cosa andremo a vedere:

- Il dato oggi:
 - tra abbondanza e complessità
 - complessità crescente, valore potenziale
- Le fondamenta del valore: architettura e qualità
- Il Data Engineer:
 - il motore invisibile
 - un ingegnere a tutti gli effetti
- Governare il dato
- I 5 messaggi chiave di oggi
- Le competenze che servono
- Strumenti, metodologie, approcci
- E ora? Cosa possiamo fare da domani
- Costruire un ecosistema del dato
- Citazioni sulla potenza dei dati



Il dato oggi: tra abbondanza e complessità 1/3.

Dati e informazioni sono spesso confusi, ma hanno significati distinti nel contesto informatico.

Dato:

- È un elemento grezzo o elementare.
- Solitamente costituito da simboli che richiedono elaborazione e contestualizzazione.
- Ad esempio, il numero “20” da solo non ha significato, ma associato all’etichetta “età”, diventa informativo.
- I dati possono essere memorizzati in file o database.
- Si suddividono in dati semplici (parole, numeri) e complessi (combinazioni di dati semplici).

Informazione:

- Risultato dell’elaborazione di più dati.
- Ha significato e rilevanza.
- Permette il processo decisionale.
- Deriva dall’interpretazione di dati e può essere utile in vari contesti.

In breve, i dati sono elementi conosciuti, mentre le informazioni derivano dall’elaborazione di tali dati.



Il dato oggi: tra abbondanza e complessità 2/3.

Viviamo in un'epoca di sovrabbondanza informativa, ma il dato utile è spesso sommerso da:

- eterogeneità tecnologica,
- scarsa qualità,
- mancanza di governance chiara.

Governare il dato è una priorità, non un'opzione.

 **Domanda chiave:** Stiamo solo accumulando dati o li stiamo trasformando in valore?



Il dato oggi: tra abbondanza e complessità 3/3.

Un mini esempio bancario:

"Pensiamo a quanti dati produciamo ogni giorno solo con bonifici, carte, app mobile e interazioni helpdesk. Ma quanti di questi dati diventano davvero utili per decidere su prodotti, rischio o marketing?"

Dati in banca oggi → eterogenei (CRM, core banking, ticketing), spesso **non integrati**, qualità **non sempre monitorata**.



Il dato oggi: complessità crescente, valore potenziale.

Ingegneria del dato = disciplina emergente.

Non è solo questione di big data, ma di:

- strutturare i processi informativi come sistemi ingegneristici,
- progettare soluzioni scalabili, robuste e manutenibili,
- governare le informazioni con criteri di qualità, affidabilità e sicurezza.

 **Il dato è un componente dell'infrastruttura informativa.**



Le fondamenta del valore: architettura e qualità 1/4.

Per creare valore servono basi solide:

- **Architetture scalabili** (Data Warehouse, Data Lake, Data Mesh...)
- **Pipeline robuste** e automatizzate
- **Data Quality** integrata nel ciclo di vita del dato

Il dato va progettato come un'infrastruttura: con metodo, visione e manutenzione.

Nessun ingegnere costruirebbe una struttura senza fondamenta: perché farlo con i dati?

Dato grezzo → Validazione → Standardizzazione → Arricchimento → Insight



Le fondamenta del valore: architettura e qualità 2/4.

◆ Data Warehouse



Cos'è: Un “magazzino dati” organizzato e strutturato.



A cosa serve: Conservare dati *puliti e certificati* per fare report, analisi e dashboard.



In banca: Contiene i dati "ufficiali", ad esempio per bilanci, indicatori, controllo di gestione.

◆ Data Lake



Cos'è: Un grande “lago” dove confluiscono dati di ogni tipo (grezzi, non ancora lavorati).



A cosa serve: Conservare tutto, anche i dati non ancora strutturati o etichettati.



In banca: Può contenere log, tracciamenti utente da app, dati testuali, audio dei contact center.



Le fondamenta del valore: architettura e qualità 3/4.

◆ Data Mesh

 **Cos'è:** Un **modello organizzativo** (non una tecnologia) che **distribuisce la gestione del dato** ai team che lo conoscono meglio.

 **A cosa serve:** Evitare che tutto il controllo stia solo nell'IT; ogni area ha un po' di responsabilità sui suoi dati.

 **In banca:** La funzione HR gestisce i dati HR, il Risk i suoi, il Marketing i suoi — ma con regole comuni.

In sintesi: abbiamo tre approcci diversi per governare i dati:

- il **Warehouse** raccoglie quelli ufficiali e già puliti,
- il **Lake** conserva anche quelli grezzi,
- la **Mesh** affida la cura del dato a chi lo conosce meglio.

"Warehouse, Lake e Mesh non si escludono: sono tre modi per trattare il dato a seconda dello stato in cui si trova e di chi lo usa. Oggi non basta più solo un magazzino: serve anche un lago per esplorare e una rete per responsabilizzare."



Pipeline robuste e automatizzate: cosa sono?



Pipeline dati = il **flusso automatico** che porta i dati:

1. dalla **fonte** (es. database, file, app)
2. fino alla **destinazione** (es. report, dashboard, AI)



Robuste

Significa che:

- **funzionano anche in caso di problemi** (ritardi, errori nei dati, mancanza di connessione)
- **si controllano da sole**: verificano la qualità, inviano alert se qualcosa non va
- hanno **log e tracciamento**: sappiamo sempre cosa è successo e quando



Esempio: se un dato da un'app mobile è vuoto o errato, la pipeline lo intercetta, lo blocca, segnala l'errore e continua a lavorare con gli altri dati.



Automatizzate

Significa che:

- non serve l'intervento umano continuo
- eseguono i passaggi in modo programmato e ripetibile
- possono essere riutilizzate su nuovi dati ogni giorno, ogni ora, ecc.



Esempio: ogni mattina alle 6:00 parte in automatico la pipeline che:

1. raccoglie i dati delle transazioni
2. li valida e li arricchisce
3. li carica in un report per il Risk Management

"Una pipeline robusta e automatizzata è come un nastro trasportatore intelligente: prende i dati dalla sorgente, li controlla, li pulisce e li consegna pronti all'uso, ogni giorno, in automatico. Senza intoppi, senza errori, e senza bisogno di metterci le mani tutte le volte."



📌 Data Quality integrata nel ciclo di vita del dato

Cosa vuol dire?

Non basta “pulire i dati alla fine”: **la qualità deve essere controllata in ogni fase, fin dall’inizio.**

🔄 Il ciclo di vita di un dato (semplificato):

1. **Creazione/Raccolta** → esempio: cliente inserisce dati nel form
2. **Trasformazione** → il dato passa nei sistemi, cambia formato, si arricchisce
3. **Storage** → finisce in un database, un warehouse, ecc.
4. **Analisi/Report** → viene letto da BI, AI, utenti
5. **Archiviazione o eliminazione** (data retention)

✓ "Qualità integrata" significa:

- Verifica **sin da quando il dato nasce** (es. controlli sui campi obbligatori, codifiche corrette)
- **Controlli automatici** nei flussi (pipeline)
- **Standard condivisi** tra IT e business
- Tracciamento di errori, anomalie, duplicazioni
- Meccanismi di **correzione e validazione**

🔔 *In parole povere:* non aspettiamo che il dato sia sporco per accorgercene. Lo controlliamo lungo tutto il percorso.

"Garantire la qualità del dato significa trattarlo come un prodotto industriale: lo controlliamo in ogni fase, dall'ingresso alla consegna. Non possiamo accorgerci di un errore solo nel report finale."



Il Data Engineer: il motore invisibile 1/2.

Il Data Engineering è il motore che trasforma dati grezzi in asset pronti all'uso.

- Non è solo tecnologia: è **processo, standard, cultura**
- Abilita analisi, AI, reporting, compliance

 Obiettivo: creare flussi di dati affidabili, riutilizzabili e accessibili.



Il Data Engineer: il motore invisibile 2/2.

Un esempio realistico di **cosa fa in banca**:

«Un Data Engineer può orchestrare il flusso che collega il database dei reclami all'analisi degli scostamenti di qualità del servizio. Oppure alimentare un report per l'audit settimanale anti-frode.»

Ad esempio una **giornata tipo**:

9:00 - controllo pipeline
10:30 - verifica anomaly su dashboard
13:00 - riunione con Risk
15:00 - deploy flusso certificazioni ESG



Il Data Engineer: un ingegnere a tutti gli effetti.

Profilo multidisciplinare che integra:

- Competenze software e sistemi distribuiti
- Conoscenze di ETL, orchestrazione, modellazione dati
- Sensibilità verso la qualità e la governance
- Capacità di comunicare con stakeholder aziendali

 È il ponte tra mondo tecnico e mondo decisionale.



Governare il dato = governare l'organizzazione.

La Data Governance non è solo per il compliance officer:

- Definisce **ruoli, responsabilità, ownership**
- Supporta trasparenza e collaborazione
- Protegge dai rischi (GDPR, sicurezza, decisioni sbagliate)

Un dato ben governato crea fiducia. E la fiducia crea valore.

Ad esempio «benefici per la Banca»:

- Meno errori nei report
- Maggiore affidabilità verso autorità
- Decisioni più rapide (es. pricing)

Collegamento a DORA: *"La governance è anche un obbligo normativo: DORA richiede tracciabilità, responsabilità, qualità e resilienza del dato."*



I 5 messaggi chiave di oggi:

1. Il dato è un **bene aziendale**, non un sottoprodotto.
2. Serve **un cambio culturale**, non solo tecnologico.
3. La qualità non si “ripulisce dopo”: va costruita a monte attraverso opportuni processi.
4. I dati creano valore **solo se condivisi e accessibili**.
5. **Governare = scegliere**: cosa raccogliere, cosa ignorare, cosa valorizzare.



Le competenze che servono.

Non bastano i tool. Servono persone che sappiano:

- progettare flussi,
- comprendere i business need,
- facilitare il dialogo tra IT e funzioni aziendali.

Figure chiave:

- Data Engineer
- Data Steward
- Data Architect
- Business Translator



Data Engineer

Costruisce i “binari” su cui viaggiano i dati.

Progetta e mantiene i flussi, automatizza i processi, garantisce che i dati arrivino dove servono, quando servono.

📌 *È il tecnico che trasforma il dato grezzo in materia prima utilizzabile.*



Data Steward

È il “custode” della qualità e coerenza dei dati.

Controlla che i dati siano corretti, completi, conformi e comprensibili. Mantiene le definizioni, i nomi, le regole.

📌 *È il garante della fiducia nel dato.*



Data Architect

Progetta l'architettura generale del sistema dati.

Decide come sono organizzati, dove stanno, come si collegano fra loro. Sceglie tecnologie e standard.

📌 *È l'ingegnere strutturale del mondo dati.*



Business Translator

Traduce i bisogni del business in soluzioni dati.

Parla con manager, marketing, finance... e con l'IT. Fa da ponte, capisce le domande aziendali e aiuta a trovare le risposte nei dati.

📌 *È il mediatore tra chi ha le domande e chi costruisce le risposte.*



Strumenti, metodologie, approcci 1/3.

Alcuni strumenti e approcci noti nel contesto tecnico:

- DataOps: CI/CD applicato ai dati
- Data Contracts & Schema Validation
- Observability & Logging nelle pipeline dati
- Standard per la modellazione: **Data Vault, Kimball, Star Schema**
- Automazione con strumenti open (Airflow, dbt, Great Expectations)



Strumenti, metodologie, approcci 2/3.

DataOps: CI/CD applicato ai dati

Porta nel mondo dei dati i principi dell'ingegneria del software:
integrazione continua, test automatici, rilascio controllato dei flussi dati.

 *Risultato: pipeline più affidabili, veloci da modificare e più facili da mantenere.*

Data Contracts & Schema Validation

Contratti tra chi fornisce e chi usa il dato: stabiliscono regole precise su **formato, contenuto, frequenza, struttura.**

 *Servono per evitare rotture nei flussi: se cambia qualcosa, lo sappiamo subito.*

Observability & Logging nelle pipeline dati

Monitoraggio continuo dei flussi: **cosa passa, quando, con quali problemi.**

 *Come avere telecamere e sensori su ogni "tubo" che trasporta dati.*



Strumenti, metodologie, approcci 3/3.

Standard per la modellazione: Data Vault, Kimball, Star Schema

Metodi collaudati per organizzare i dati nei database:

- **Kimball e Star Schema** → orientati alla reportistica
- **Data Vault** → pensato per flessibilità, audit, tracciabilità

📌 Scelti in base agli obiettivi (stabilità, velocità, compliance).

Automazione con strumenti open

Strumenti gratuiti e potenti per gestire pipeline:

- **Airflow**: orchestration dei flussi
- **dbt**: trasformazione dei dati e versionamento
- **Great Expectations**: test automatici sulla qualità del dato

📌 Riducono l'intervento manuale e aumentano il controllo.



E ora? Cosa possiamo fare da domani:



- Valutare la maturità dei propri dati (data maturity assessment)
- Identificare **silos e colli di bottiglia**
- Introdurre un linguaggio condiviso sui dati
- Investire in **formazione e team cross-funzionali**
- Sperimentare piccoli progetti data-driven

Cosa può fare un ingegnere, da domani:



- Portare **rigore metodologico** nella gestione dei dati
- Inserire logiche di **documentazione e versionamento**
- Promuovere pratiche di **manutenzione del dato** in azienda o nei progetti
- Fare da **facilitatore tecnico** tra business e IT
- Valorizzare le proprie competenze con certificazioni e formazione mirata



Costruire un ecosistema del dato.

Il valore dei dati non nasce da un reparto, ma da una **rete**:

- Tecnologia + Persone + Processi
- Collaborazione tra IT, governance, business
- Cultura **dell'apprendimento continuo**



“Data-driven” è un percorso, non un’etichetta.

Essere **data-driven** non significa solo **avere tanti dati o strumenti avanzati**, ma **saperli usare per decidere, migliorare, innovare**.

È un **cambiamento culturale e organizzativo**, che richiede:

- fiducia nei dati,
- qualità e governance,
- collaborazione tra funzioni,
- competenze diffuse.



Non basta dirsi “data-driven”: bisogna dimostrarlo ogni giorno, nelle scelte e nei processi.



Citazioni sulla potenza dei dati:

- **"Without data, you're just another person with an opinion." – W. Edwards Deming**
(Senza dati, sei solo un'altra persona con un'opinione.)
- **"In God we trust, all others bring data." – W. Edwards Deming**
(In Dio ci fidiamo, gli altri portano i dati.)
- **"Data is the new oil." – Clive Humby**
(I dati sono il nuovo petrolio.)
- **"The goal is to turn data into information, and information into insight." – Carly Fiorina**
(L'obiettivo è trasformare i dati in informazioni, e le informazioni in intuizioni.)



GRAZIE PER L'ATTENZIONE

"Il dato non è solo un asset tecnico, è un acceleratore decisionale."

Ing. Francesco Micillo
francescomicillo@libero.it