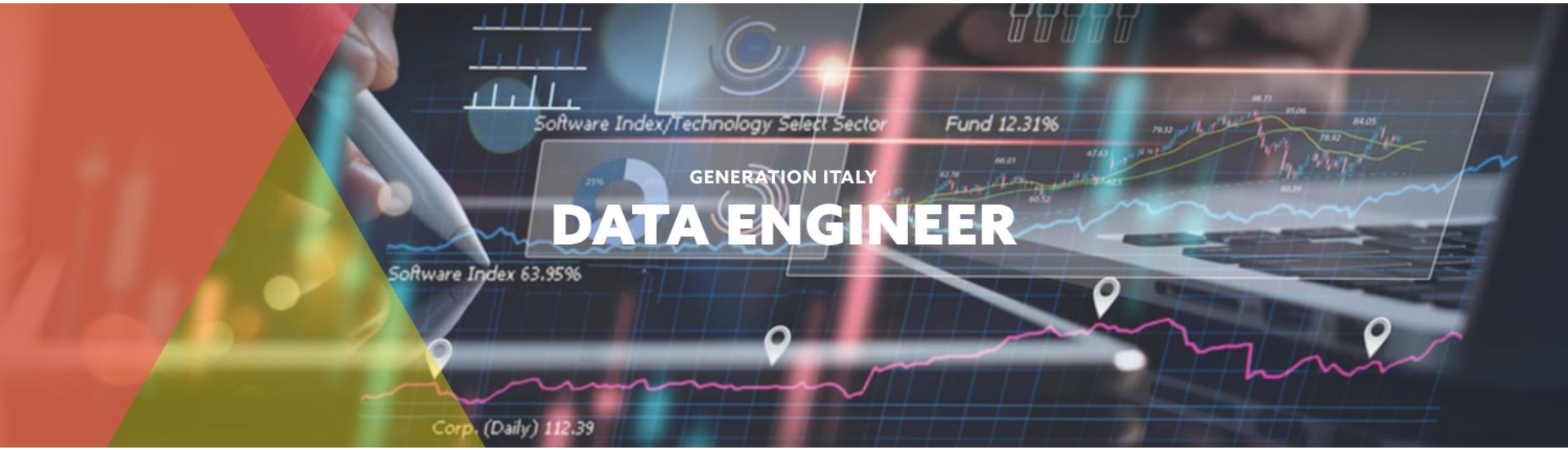
ACCADEMIA DEL LEVANTE
WWW.ACCADEMIADELLEVANTE.ORG

Generation ITALY

IASEM
Istituto Alti Studi Euro Mediterranei

# Data Engineer




## Concetti introduttivi

Ing. Maurizio Maggiora

# Cos’è la data science?

Question Time

# Cosa è la Data Science?

- La Data Science è l'intersezione tra statistica e informatica
- Si tratta di estrarre conoscenza o insight dai dati per prendere decisioni informate
- L'analisi di dati è sempre esistita, si parla di Data Science quando è richiesta una **competenza tecnica** per estrapolare e manipolare i dati e **competenza matematica e statistica** per analizzare grandi volumi di dati e fare analisi predittive

# I numeri del mercato del lavoro

- La data science è stata nominata il lavoro in più rapida crescita nel 2017 da LinkedIn e nel 2018 Glassdoor ha classificato il data scientist come il miglior lavoro negli Stati Uniti. Inoltre, un recente studio di PriceWaterhouseCoopers afferma: "I migliori lavori in questo momento in America includono titoli come data scientist, data engineer e business analyst"
- Crescita del 650% annuo dal 2012
- Stipendio medio USA 110k
- La crescita del settore dei big data è stata confermata anche dalla Harvard Business Review che ha definito il data scientist come la professione più “sexy” del 2021. Anche la classifica di Glassdoor sui “50 best jobs in America” ha messo il data scientist al terzo posto nel 2022 e al primo posto tra il 2016 e il 2019. Secondo il Bureau Labor Statistics degli Stati Uniti i posti di lavoro in questo settore sono destinati ad aumentare dell’11% entro il prossimo anno

# E in Italia?

- L'azienda Experis, nel suo report "Tech cities" 2022, ha definito il data scientist come il secondo profilo più richiesto in Italia (17%) dopo il Java developer (46%). Le offerte di lavoro si concentrano soprattutto nel nord Italia, in particolare a Milano con il 53%; Roma in seconda posizione con il 20,4%.
- Scegliere questo campo professionale serve anche ad evitare il grande problema – non solo italiano – del lavoro sottopagato. La retribuzione annua lorda dei data scientist parte (in media) da un minimo di 27 K€ fino ad arrivare a 40 K€.
  Gli stipendi più alti sono in Lombardia e Piemonte dove il guadagno media annuo è rispettivamente di 40 K€ e 36 K€.

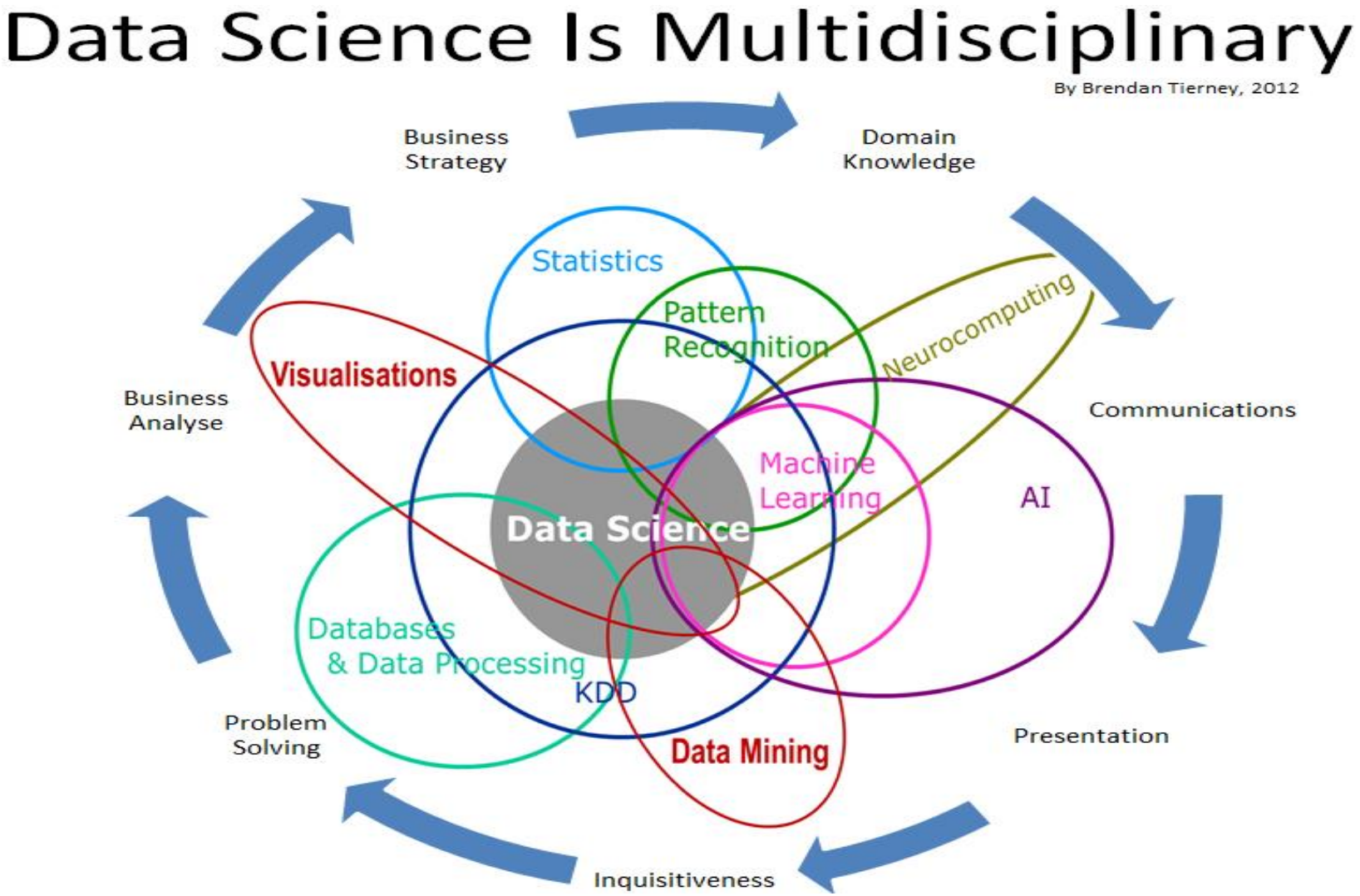
# Sì ma… che lavoro è?

- Si tratta di un professionista con competenze che vanno dall'informatica alla statistica alla matematica.
- L'obiettivo principale è l'organizzazione, l'analisi e l'interpretazione di una grande quantità di dati, il tutto supportato dall'utilizzo di software progettati ad hoc.

# Competenze necessarie

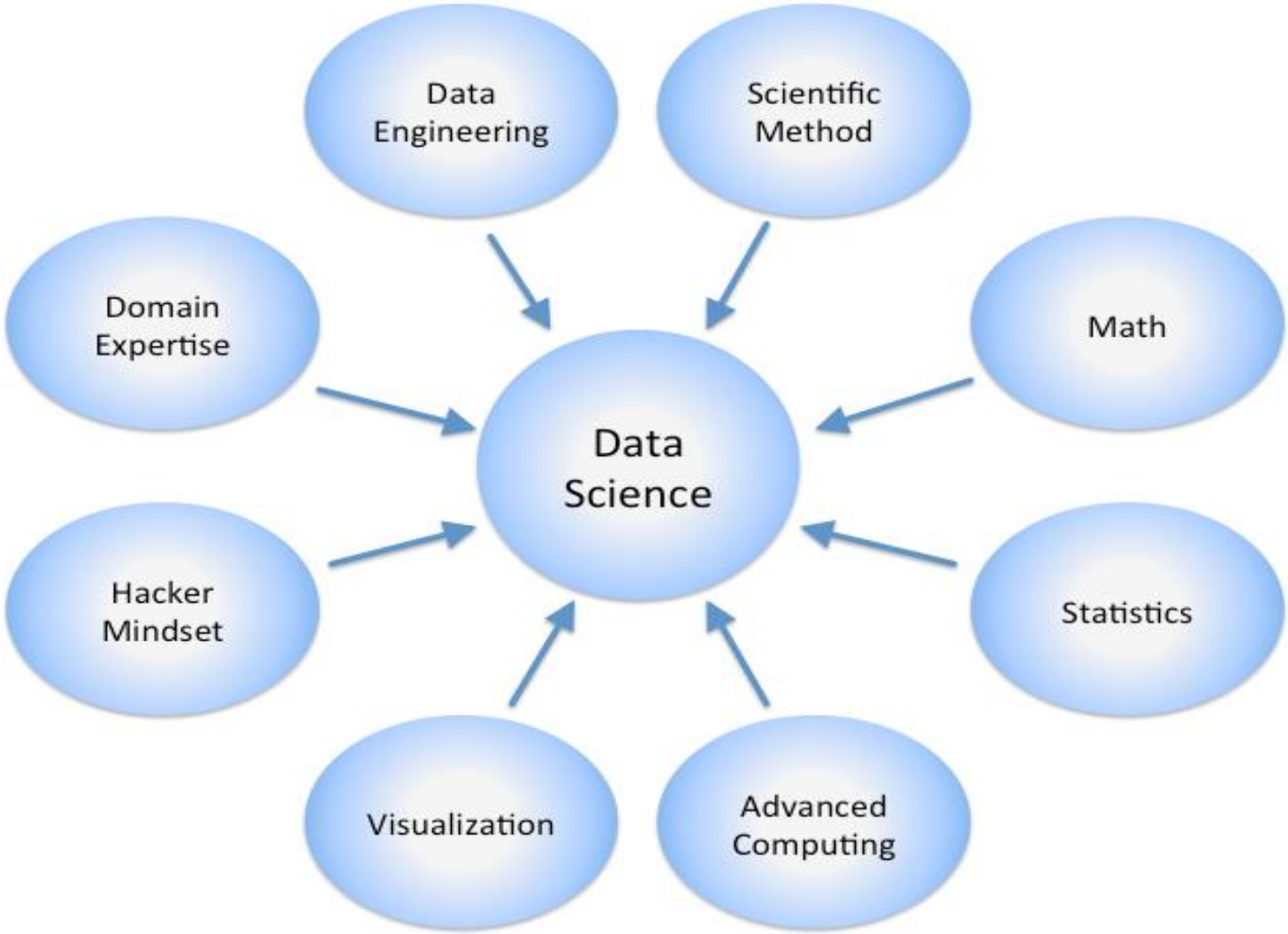
- Il Data Scientist deve coniugare le skill tecniche e l'intuizione per organizzare grandi set di dati e rispondere a domande complesse, elaborando report che aiutano i top manager a definire le strategie.
- Deve saper navigare tra **dati strutturati** (organizzati per categorie, come i dati di vendita) e **dati non strutturati** (più difficili da classificare in modo automatizzato, come i commenti sui social media) conducendo analisi quantitative e qualitative.
- Sono necessarie soft skill per la comprensione del business.

# Competenze necessarie



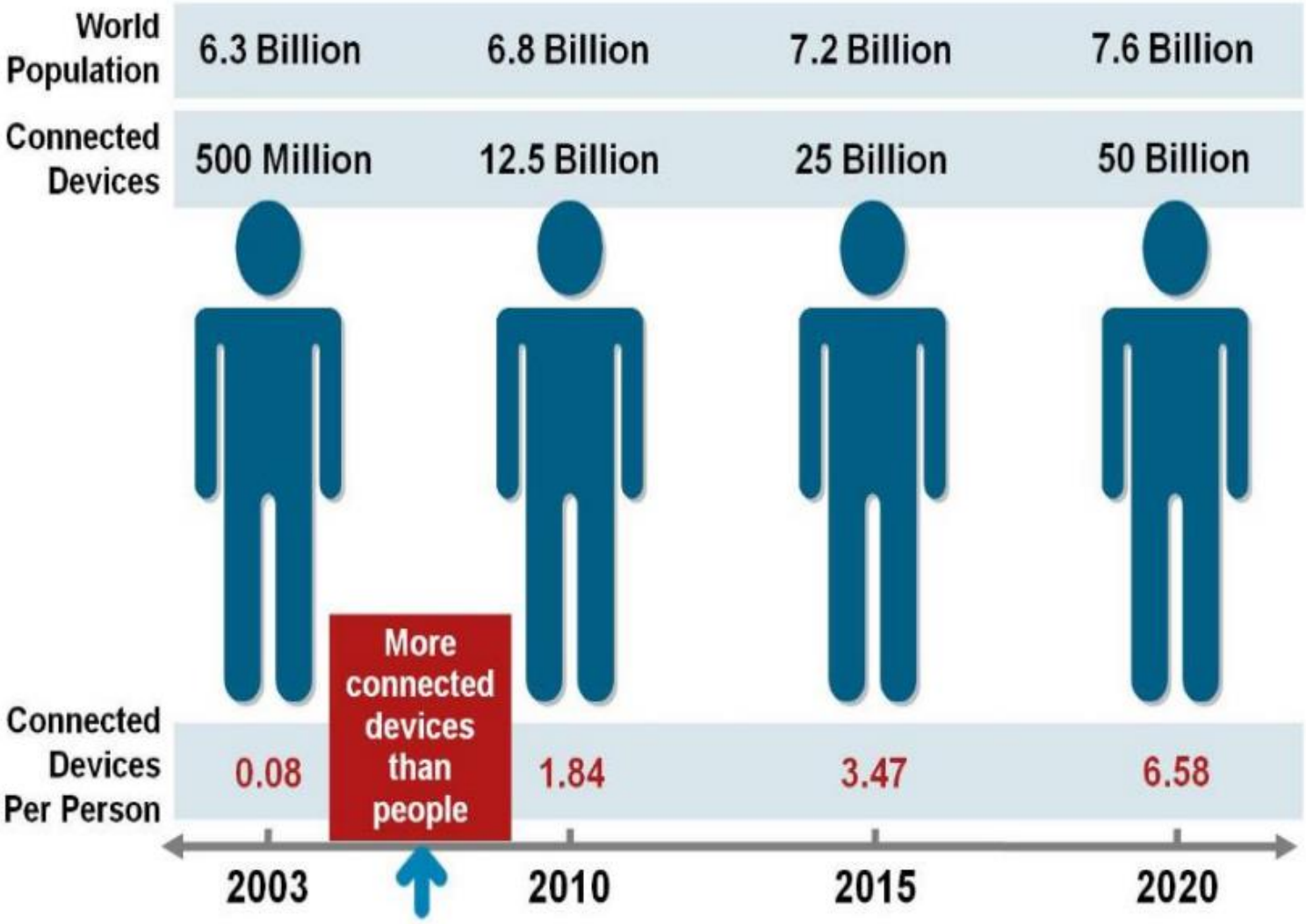
ACCADEMIA DEL LEVANTE
WWW.ACCADEMIADELLEVANTE.ORG

# Competenze necessarie



# Perché ora?

Question Time



# Ruoli e differenze

## Data Analyst

Most entry-level professionals interested in getting into a data-related job start off as Data analysts. All you need is a good statistical knowledge. Strong technical skills would be a plus and can give you an edge over most other applicants. Other than this, companies expect you to understand data handling, modeling and reporting techniques along with a strong understanding of the business.

# Ruoli e differenze

**Data Engineer**

Data Engineer gathers a good amount of experience as a Data Analyst. A Data Engineer needs to have a strong technical background with the ability to create and integrate APIs. They also need to understand data pipelining and performance optimization.

# Ruoli e differenze

## Data Scientist

Data Scientist is the one who analyses and interpret complex digital data. While there are several ways to get into a data scientist's role, the most seamless one is by acquiring enough experience and learning the various data scientist skills. These skills include advanced statistical analyses, a complete understanding of machine learning, data conditioning etc.

# Ruoli e differenze

| Data Analyst | Data Engineer | Data Scientist |
|---|---|---|
| Pre-processing and data gathering | Develop, test & maintain architectures | Responsible for developing Operational Models |
| Emphasis on representing data via reporting and visualization | Understand programming and its complexity | Carry out data analytics and optimization using machine learning & deep learning |
| Responsible for statistical analysis & data interpretation | Deploy ML & statistical models | Involved in strategic planning for data analytics |
| Ensures data acquisition & maintenance | Building pipelines for various ETL operations | Integrate data & perform ad-hoc analysis |
| Optimize Statistical Efficiency & Quality | Ensures data accuracy and flexibility | Fill in the gap between the stakeholders and customer |

ML = Machine Learning
ETL = Extract, Transform, Load

# La base dati: i Big data

I Big Data sono un'enorme quantità di dati generati da qualunque dispositivo connesso ad Internet, una vera e propria raccolta di dati così estesa in termini di volume, velocità e varietà, da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza.

Big data è un termine usato per descrivere gli enormi volumi di dati digitali generati, raccolti ed elaborati. Il termine big data descrive i dati che si muovono troppo velocemente, sono semplicemente troppo grandi o troppo complessi per essere archiviati, elaborati o analizzati con le tradizionali applicazioni di archiviazione e analisi dei dati.

Alcuni esempi di big data includono i dati generati dai post sugli account dei social media, come Facebook e Twitter, e le valutazioni date ai prodotti sui siti di e-commerce come Amazon.

# Le 3 V… più una

- **VOLUME**
  Grande quantità di dati generati
- **VARIETA'**
  Differenti tipologie di dati
- **VELOCITA'**
  I dati viaggiano in rete ad una velocità elevatissima generandone sempre di nuovi
- **VERIDICITA'**
  Qualità e l'integrità dei dati, aspetto fondamentale per un'analisi utile ed affidabile

# Le 3 V… più una: Volume

Il volume descrive la quantità di dati trasportati e archiviati. Secondo gli esperti di International Data Corporation (IDC), scoprire modi per elaborare la crescente quantità di dati generati ogni giorno è una sfida. Prevedono che il volume dei dati aumenterà a un tasso di crescita annuo composto del 23% nei prossimi cinque anni. Mentre i tradizionali sistemi di archiviazione dei dati possono, in teoria, gestire grandi quantità di dati, stanno lottando per tenere il passo con le richieste di grandi volumi di big data.

# Le 3 V… più una: Varietà

Varietà descrive le molte forme che i dati possono assumere, la maggior parte delle quali raramente sono pronte per l'elaborazione e l'analisi. Un contributo significativo ai big data sono i dati non strutturati, come video, immagini e documenti di testo, che si stima rappresentino dall'80 al 90% dei dati mondiali. Questi formati sono troppo complessi per le tradizionali architetture di archiviazione di data warehouse. I dati non strutturati che costituiscono una parte significativa dei big data non rientrano nelle righe e nelle colonne del tradizionale sistema di archiviazione dei dati relazionali.

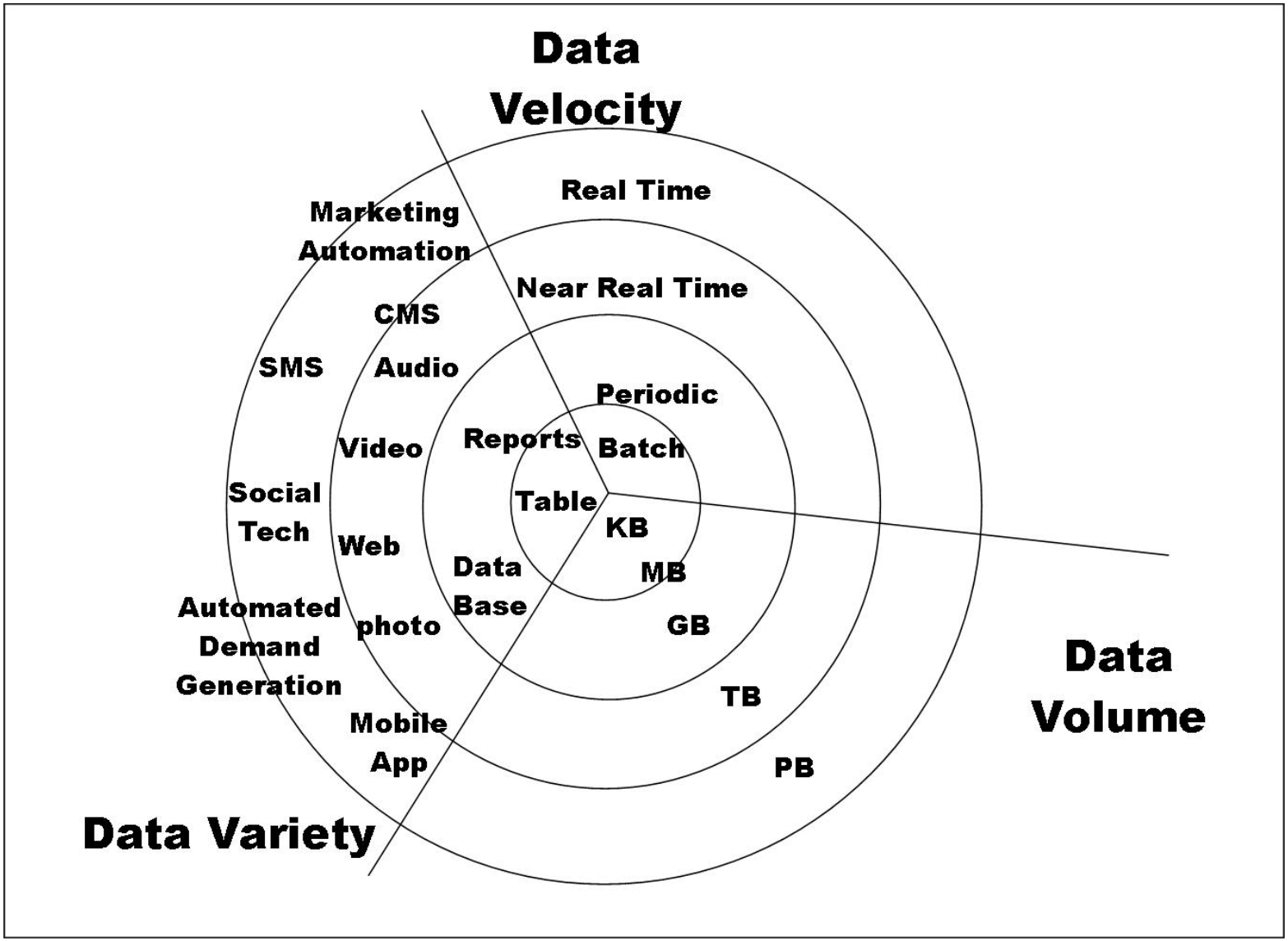
# Le 3 V… più una: Velocità

La velocità descrive la velocità con cui questi dati vengono generati. Ad esempio, i dati generati dalla Borsa di New York da un miliardo di azioni vendute non possono essere archiviati per un'analisi successiva, ma devono essere analizzati. L'infrastruttura dati deve rispondere istantaneamente alle richieste delle applicazioni che accedono e trasmettono i dati. I big data scalano istantaneamente e la ricerca spesso deve avvenire in tempo reale.

# Le 3 V… più una: Veridicità

La veridicità è il processo per impedire che dati imprecisi rovinino i tuoi set di dati. Ad esempio, quando le persone si registrano per un account online, spesso utilizzano false informazioni di contatto. Gran parte di queste informazioni imprecise devono essere "cancellate" dai dati prima dell'uso nell'analisi. Una maggiore veridicità nella raccolta dei dati può ridurre la quantità di pulizia dei dati necessaria.

# Le 3 V… più una



# La registrazione del dato

Prerequisito per poter fare analisi dati: **l'esistenza del dato**.

Se il dato non è registrato elettronicamente, non esiste (a meno che non venga poi trasformato in dato elettronico). Non a caso si parla di DATABASE (base dati) .

Esempi di dato non registrato:

- pagamento a nero
- dati registrati a penna su un foglio
- azioni derivanti da conversazioni telefoniche o di persona non tracciate

# Dataset di grandi dimensioni

- Le aziende non devono necessariamente generare i propri Big Data.
- Esistono fonti di set di dati (**dataset**) gratuiti disponibili, pronti per essere utilizzati e analizzati.

# Quali sono le sfide dei Big Data?

- Le stime dei Big Data di IBM concludono che «ogni giorno creiamo 2,5 quintillioni ($10^{30}$) di byte di dati».
- Esempi di dati raccolti dai sensori:
  - Sensori in un'auto autonoma: 4.000 GB di dati al giorno
  - Airbus A380 Engine su un volo da Londra a Singapore: 1 PB di dati
  - Sensori di sicurezza nelle operazioni minerarie: 2,4 TB di dati al minuto
  - Sensori in una casa intelligente e connessa: 1 GB di dati a settimana

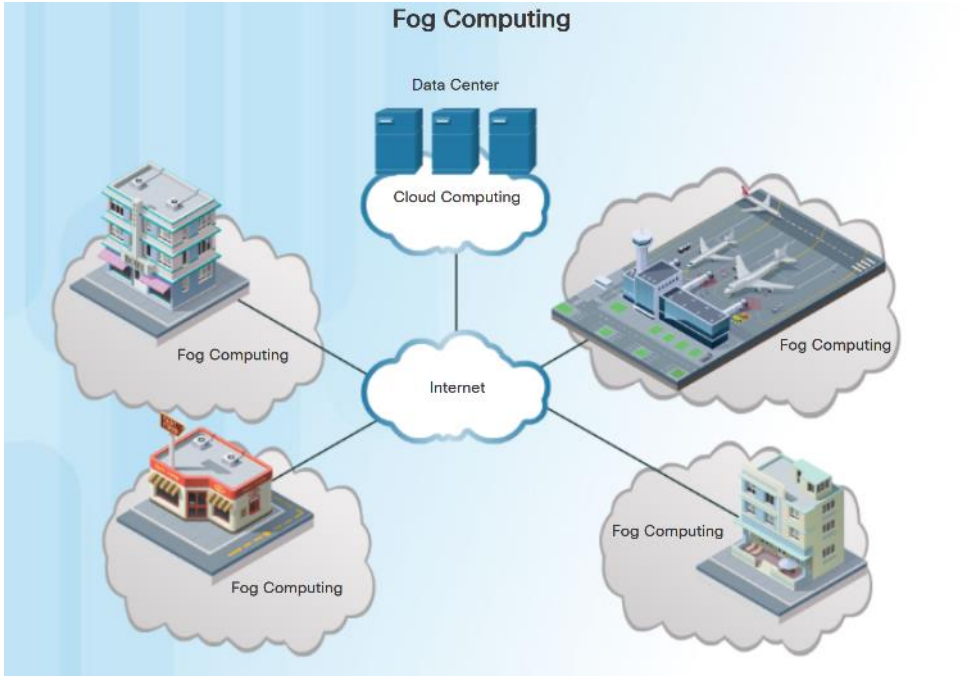
# Quali sono le sfide dei Big Data?

Problemi relativi all'archiviazione dei Big Data

- Gestione
- Sicurezza
- Ridondanza
- Analisi dei dati
- Accesso

# Dove sono archiviati i Big Data?

- I Big Data sono in genere archiviati su più server, nei data center.
- Il Fog Computing utilizza i client dell'utente finale o dispositivi «edge» per eseguire una notevole quantità di pre-elaborazione e archiviazione.
  - I dati di tale analisi pre-elaborata possono essere inseriti nei sistemi delle aziende per modificare i processi, se necessario.
  - Le comunicazioni da e verso i server e i dispositivi sono più veloci e richiedono una larghezza di banda inferiore rispetto all'utilizzo costante del cloud.



# Il cloud e il cloud computing

- Il cloud è una raccolta di data center o gruppi di server connessi.
- I servizi cloud per gli individui includono:
  - Archiviazione di dati, come immagini, musica, film ed e-mail.
  - Accesso a molte applicazioni invece di scaricarle sul dispositivo locale.
  - Accesso a dati e applicazioni ovunque, in qualsiasi momento e su qualsiasi dispositivo.
- I servizi cloud per le aziende includono:
  - Accesso ai dati organizzativi ovunque e in qualsiasi momento.
  - Semplificazione delle operazioni IT di un'organizzazione.
  - Eliminazione o riduzione della necessità di apparati IT onsite e relativa manutenzione e gestione.
  - Riduzione dei costi per gli apparati, l'energia, i requisiti di impianto fisico e la necessità di formazione del personale.

# La registrazione del dato

Dato GREZZO → Registrazione → Dati presenti in un database → Eventuale collegamento di database diversi → Estrapolazione dati intelligente → Visualizzazione, report e analisi

- Il **data mining** è il processo di trasformazione dei dati grezzi in informazioni significative.
- I dati estratti devono essere analizzati e presentati a manager e decisori.

# Indica il numero di dispositivi in tuo possesso che possono essere fonte di trasmissione di dati

Question Time

# Tipo di dati

Tutti i dati devono specificare il «tipo di dati» che indica alle applicazioni come trattarli. Le operazioni eseguite sono definite dal tipo di dati. L'identificazione dei tipi di dati è utile nell'analisi perché potrebbe essere necessario raggruppare i dati, ordinare i dati o eseguire calcoli sui dati.

I dati devono essere raggruppati in base al tipo specificato per eseguire le operazioni richieste.

# Tipo di dati

- Strutturati

I dati strutturati costituiscono circa il 10% -20% dei dati generati e hanno tipi di dati e modelli chiaramente definiti che li rendono facilmente archiviabili e organizzati in colonne e righe. Questa organizzazione semplifica la ricerca e l'analisi dei dati strutturati. Le fonti di dati strutturati includono i registri delle vendite, i sistemi di prenotazione delle compagnie aeree e il controllo dell'inventario. I dati strutturati vengono solitamente archiviati in **database relazionali** come i database SQL (Structured Query Language) o in **fogli di calcolo** come Microsoft Excel.

# Tipo di dati

- Non strutturati

I dati non strutturati costituiscono la maggior parte dei dati generati, circa l'80%, e non possono essere organizzati in righe e colonne. Ciò rende difficile la ricerca, la gestione e l'analisi dei dati non strutturati. Le fonti di dati non strutturati includono immagini, PDF, dati dei sensori e post sui social media. I dati non strutturati vengono solitamente archiviati in un **database non relazionale** noto anche come database NoSQL.

# Tipo di dati

- Volontari (social media, consenso a policy website)
- Osservati (location, siti web visitati)
- Dedotti (dedotti dalle azioni – stile di vita, gusti e preferenze)