UNIVERSITÀ DEGLI STUDI DI TERAMO

Big Data Analytics

Introduzione ai dati

Prof.ssa Romina Eramo
Università degli Studi di Teramo
Dipartimento di Scienze della Comunicazione
reramo@unite.it

Viviamo nella Data-Driven Economy

» Viviamo in una società sempre più guidata dai dati: ogni interazione digitale genera informazioni
  - I dati sono il "petrolio del XXI secolo"
» Social e macchine generano enormi quantità di informazioni
» Le aziende li usano per ottenere **vantaggi competitivi**
» Gestire bene i dati interni significa:
  - Maggiore efficienza dei processi
  - Aumento della produttività
  - Analisi dei punti di forza/debolezza
  - Valutazione dei rischi e strategie di prevenzione

Big Data: il panorama odierno

» I Big Data rappresentano **grandi volumi di dati** prodotti ad alta velocità e in diversi formati
  - Fonti: social network, sensori IoT, e-commerce, log di sistema, dispositivi mobili, sistemi industriali
  - Dati complessi che non possono essere facilmente gestiti o analizzati con gli strumenti di elaborazione dei dati tradizionali, in particolare i fogli di calcolo

» Le tecnologie Big Data permettono di **estrarre valore**, migliorare decisioni, ottimizzare processi e prevedere comportamenti

» Competenze, infrastrutture e algoritmi sono oggi elementi chiave per **trasformare i dati in vantaggio competitivo**

Nuove professioni: il Data Scientist

» Il **Data Scientist** unisce competenze informatiche, statistiche e analitiche

» È oggi tra i lavori più richiesti e meglio retribuiti

» Supporta decisioni strategiche e innovazione aziendale

» Ha un impatto su performance, trasparenza e competitività

I dati in azienda: attori e ruoli

» I manager usano dati sintetici per decisioni strategiche/tattiche

» Il personale operativo necessita di dati dettagliati

» Il data scientist elabora e organizza i dati per entrambi

- Figura centrale nell'estrazione di valore da dati e Big Data

» **Competenze del Data Scientist**

- Conoscenza di architetture e tecnologie per l'elaborazione dati
- Tecniche descrittive e predittive
- Creazione di piattaforme analitiche per management e operativi
- Integrazione di fonti interne ed esterne

Panoramica dei dati aziendali

» I dati sono un asset strategico

» Analizzare le fonti, i supporti tecnologici e le strutture

» La gestione efficace dei dati aumenta l'efficienza e la competitività

» **Fonti dei dati aziendali**

- **Fonti interne**: operazionali, data warehouse, data mart
- **Fonti esterne**: clienti, fornitori, social, web
- Le fonti variano in base al settore

Fonti interne

Fonti operazionali (esempi industriali)	Fonti operazionali (altri settori)
Gestione produzione: materie prime, consumi, output **Acquisti e magazzino**: ordini, movimenti **Ordini e consegne**: logistica **Contabilità**: fatture, cassa, banca **HR**: stipendi, premi, obiettivi **CRM**: anagrafiche clienti, marketing	**Banche**: sportelli, strumenti finanziari, rischio, prestiti **GDO**: scontrini, fidelity card, promozioni, vendite Volumi elevati in settori come finanza, industria e retail

Fonti interne

» Perché non analizzare direttamente i sistemi operazionali?
- Software eterogenei, prodotti da vendor diversi, con strutture dati non uniformi
- Dati replicati e incoerenti tra sistemi (es. anagrafiche clienti, fornitori…)
- Strutture dati OLTP: ottimizzate per le transazioni, non per l'analisi
- Difficoltà nell'estrazione: dati normalizzati, molti join necessari
- Scarsa profondità storica: difficile ricostruire il passato

Data Warehouse e Data Mart

Database progettati per contenere dati integrati, coerenti e certificati
Raccolta centralizzata di dati da più fonti operative
Permettono analisi storiche e trasversali
Base per la Business Intelligence (BI): modelli, strumenti, processi
Supportano decisioni* manageriali basate su dati affidabili

* I sistemi di Business Intelligence sono anche definiti come Decision Support System (DSS).

Fonti interne

Basi dati ad hoc

Sono create per esigenze analitiche specifiche
Contengono rielaborazioni di dati operazionali o provenienti dal data warehouse
Spesso sono di “proprietà” del singolo analista
Possono contenere dati molto preziosi, utili anche ad altri reparti
Spesso non sono integrate nei flussi ufficiali, ma rappresentano una risorsa strategica

Fonti esterne

Fonti esterne di dati

I dati esterni completano o integrano quelli interni
Esempi: anagrafica ISTAT, social media, blog, forum
Alcune analisi si basano solo su dati esterni (es. sentiment analysis)
Utili per comprendere opinioni, trend, percezioni su temi o aziende

» Problemi legati ai dati esterni

- Qualità dei dati non sempre verificabile o garantita
- L'azienda non controlla accuratezza, completezza e consistenza (qualità dei dati)
- Verifica complessa o impossibile
- Serve attenzione nella selezione e nel trattamento delle fonti

Qualità dei dati: caratteristiche

- Completezza
  - Presenza di tutti i dati necessari per descrivere un'entità, una transazione o un evento.
  - *Esempio*: un'anagrafica è incompleta se mancano campi essenziali (es. codice fiscale, indirizzo).
- Consistenza
  - Assenza di contraddizioni nei dati, garantendo coerenza interna tra valori.
  - *Esempio*: in una banca, il saldo di fine mese deve coincidere con quello risultante da saldi precedenti e movimenti.
- Accuratezza
  - I dati devono essere corretti e conformi alla realtà.
  - *Esempio*: un indirizzo o un importo errato riduce la validità dell'informazione.
- Assenza di duplicazione
  - Ogni dato (campo, record, tabella) deve essere unico all'interno o tra i sistemi.
  - *Problema*: la duplicazione può causare disallineamenti, errori e doppia manutenzione.
- Integrità
  - I dati devono rispettare vincoli definiti a livello di database (chiavi primarie/esterne, tipi di dato, vincoli logici).
  - *Esempio*: nessun valore non valido in una colonna vincolata a un elenco predefinito

I tipi di struttura dei dati

» **Dati strutturati**: organizzati in formato tabellare (es. CSV, XML, JSON)

→ Attributi chiari, facilmente gestibili in un database relazionale

» **Dati non strutturati**: privi di schema definito (es. testo libero, immagini, video, audio)

→ Non gestibili efficacemente in forma tabellare

» **Dati semi-strutturati**: parzialmente organizzati

→ Es. file Word o immagini con metadati (titolo, autore, GPS, ecc.)

→ Contengono sia struttura (metadati) che contenuto libero (testo, pixel…)

Trattamento dei dati non strutturati

» I dati non strutturati richiedono **trasformazioni** per essere analizzati

» **Esempio: analisi del sentiment** su tweet

- Preprocessing del testo: pulizia e normalizzazione
- Creazione della **document-term matrix**:
  - » Righe = tweet
  - » Colonne = parole
  - » Celle = occorrenze delle parole
- Questa struttura consente l'uso di tecniche predittive e statistiche

La provenienza dei dati

» I dati possono essere classificati in base alla loro origine:
  - **Dati generati dalle persone**: social network, e-commerce, operazioni di data entry manuale
  - **Dati generati dalle macchine**: sensori, strumenti di misurazione, log di sistemi, lettori di codici a barre, sistemi di calcolo

» I dati generati automaticamente tendono ad avere **una qualità più stabile**, ma possono essere soggetti ad **anomalie in caso di malfunzionamenti** (es. sensore guasto)

Attori aziendali e dati

Manager	Personale esecutivo	Data Scientist
• Obiettivi: decisioni strategiche e tattiche • Focus su analisi predittive per anticipare trend e scenari di business	• Si occupa dell'operatività quotidiana • Richiede **dati dettagliati e aggiornati in tempo reale** • Le analisi sono granulari, utili per migliorare efficienza dei processi	• Analisi avanzata e valorizzazione del dato • Competenze principali: **SQL, Hadoop, Spark, NoSQL**, **Python/R**, tecniche di **machine learning** e visualizzazione, **Qualità del dato, business e dominio settoriale** • **Team multidisciplinari**

Archiviazione dei dati

File system e archivi flat

- File **CSV, Excel, JSON, XML, TXT, Parquet**, ecc.
- Possono essere locali o su cloud (Google Drive, Dropbox, S3…)
- Tipico uso: scambio dati, report estemporanei, archiviazione storica

Fogli di calcolo (es. Excel, Google Sheets)

- Ancora molto usati per analisi preliminari o personali
- Limiti in scalabilità, tracciabilità e sicurezza
- A volte usati come *basi dati ad hoc*

Archiviazione dei dati (2)

Database relazionali (RDBMS)

- Struttura: tabelle, righe, colonne (secondo la teoria relazionale)
- Esempi: **Oracle, MySQL, PostgreSQL, SQL Server**
- Vantaggi: coerenza, integrità referenziale, accesso tramite SQL
- Tipico uso: sistemi transazionali, ERP, CRM, contabilità

Database non relazionali (NoSQL)

- Modelli: **documenti (MongoDB), key-value (Redis), grafi (Neo4j), colonne (Cassandra)**
- Vantaggi: alta scalabilità, flessibilità, adatti a dati non strutturati
- Tipico uso: Big Data, social media, IoT, log, contenuti dinamici

Archiviazione dei dati (3)

Data warehouse

- Supporti ottimizzati per analisi e reporting storico
- Esempi: **Amazon Redshift, Snowflake, Teradata, Google BigQuery**
- Struttura: dati denormalizzati, organizzati per facilitare interrogazioni complesse
- Tipico uso: Business Intelligence, dashboard, analisi strategica

Data mart

- Versione più piccola e specifica di un data warehouse
- Focus su un singolo reparto o ambito aziendale (es. marketing, vendite)
- Vantaggi: più semplici da gestire, mirati

Archiviazione dei dati (3)

Data lake

- Architetture per raccogliere grandi volumi di dati grezzi
- Adatti a contenuti strutturati, semi-strutturati e non strutturati
- Tipico uso: machine learning, AI, analisi avanzata
- Tecnologie: Hadoop, Azure Data Lake, AWS Lake Formation

Capitolo 1 - Big Data Anlytics, A. Rezzani