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

I dati come risorsa strategica per organizzazioni e società

“I dati sono il petrolio del XXI secolo”

EFFICIENZA

Migliore gestione dei processi e delle risorse.

PRODUTTIVITÀ

Analisi dei punti di forza e di debolezza.

PREVENZIONE

Valutazione dei rischi e scelte più consapevoli.

Big Data: il panorama odierno

Volume, velocità e varietà dei dati cambiano il modo di analizzare

FONTI

Social network, IoT (Internet of Things), e-commerce, log, dispositivi mobili e sistemi industriali.

SFIDA

Dati complessi, difficili da trattare con soli fogli di calcolo e strumenti tradizionali.

TECNOLOGIE

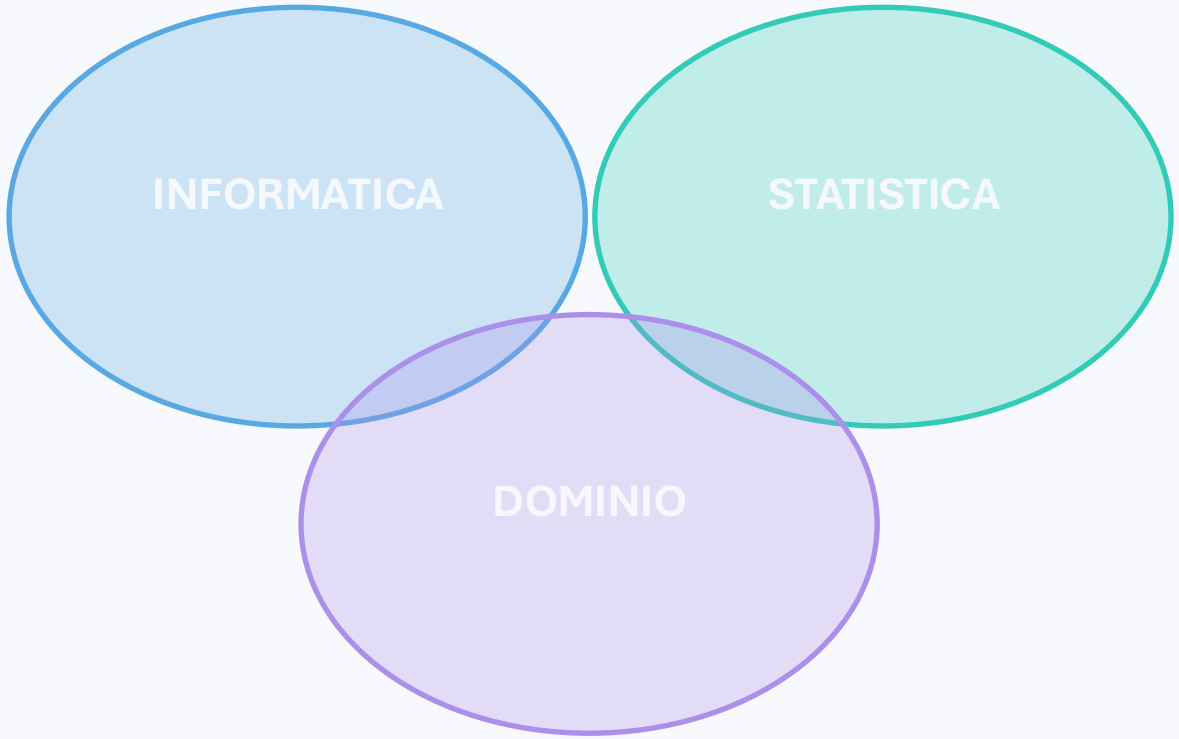
Competenze, infrastrutture e algoritmi per raccogliere ed elaborare grandi dataset.

VALORE

Migliorare decisioni, ottimizzare processi e prevedere comportamenti.

Nuove professioni: il Data Scientist

Una figura all’intersezione tra tecnologia, statistica e business



Analisi dei dati • Decisioni strategiche • Innovazione

I dati in azienda: attori e ruoli

Esigenze informative diverse, un patrimonio comune

MANAGER

Sintesi, KPI (Key Performance Indicator) e dati aggregati per decisioni strategiche e tattiche.

PERSONALE OPERATIVO

Dati dettagliati e aggiornati per le attività quotidiane.

DATA SCIENTIST

Organizza, integra e analizza dati per supportare entrambi.

Competenze: architetture dati · analisi descrittiva e predittiva · integrazione delle fonti · piattaforme analitiche

Panoramica dei dati aziendali

Fonti, tecnologie e strutture da conoscere

FONTI INTERNE

- Dati operazionali
- Data warehouse e data mart
- CRM (Customer Relationship Management)
- Basi dati ad hoc

FONTI ESTERNE

- Clienti e fornitori
- Social media e web
- Dati statistici pubblici
- Fonti specifiche del settore

Fonti interne

I dati operazionali dipendono dal settore

AZIENDA INDUSTRIALE

Produzione: materie prime e output
Acquisti e magazzino: ordini e movimenti
Logistica: ordini e consegne
Contabilità e HR (Human Resources), CRM

BANCHE E GRANDE DISTRIBUZIONE

Banche: sportelli, prestiti, strumenti finanziari e rischio

Retail: scontrini, fidelity card, promozioni, vendite

In settori quali finanza, industria e retail i dati operazionali possono raggiungere volumi molto elevati.

Dai sistemi operazionali ai Data Warehouse

Perché non analizzare direttamente i dati transazionali?

LIMITI DEI SISTEMI OLTP

OLTP = Online Transaction Processing
- Software e formati eterogenei
- Repliche e incongruenze
- Dati normalizzati: molti JOIN
- Storico spesso limitato
- Ottimizzazione per transazioni

DATA WAREHOUSE E DATA MART

- Dati integrati e coerenti
- Raccolta da fonti diverse
- Analisi storiche e trasversali
- Base della BI (Business Intelligence)
- Supporto alle decisioni

Sistemi operazionali → Integrazione dei dati → Analisi / Business Intelligence

Fonti interne: basi dati ad hoc

Un patrimonio analitico spesso poco visibile

BASI DATI AD HOC

Create per esigenze analitiche specifiche, spesso da singoli analisti.
Possono rielaborare dati operazionali o del data warehouse.
Il loro contenuto può essere prezioso anche per altri reparti, ma non sempre entra nei flussi ufficiali.

Fonti esterne

Ampliare la conoscenza, valutando attentamente la qualità

POTENZIALITÀ

- Integrazione dei dati interni
- ISTAT (Istituto Nazionale di Statistica) e altre fonti pubbliche
- Social media, blog e forum
- Sentiment analysis e analisi dei trend

CRITICITÀ

- Accuratezza non sempre verificabile
- Incompletezza e incoerenza
- Controllo limitato sulla fonte
- Necessità di selezione e pulizia

Qualità dei dati: caratteristiche

La dimensione del dataset non garantisce la qualità

COMPLETEZZA

Campi e osservazioni necessari presenti.

CONSISTENZA

Nessuna contraddizione tra valori.

ACCURATEZZA

Corrispondenza con la realtà.

NON DUPLICAZIONE

Nessuna replica incoerente.

INTEGRITÀ

Rispetto dei vincoli definiti.

I tipi di supporto

Le tecnologie di archiviazione dipendono dagli obiettivi

RDBMS E SQL

RDBMS = Relational Database Management System
SQL = Structured Query Language
File CSV = Comma-Separated Values; TSV = Tab-Separated Values.

SMP

SMP = Symmetric Multiprocessing
Memoria e I/O (Input/Output) condivisi.
Possibili colli di bottiglia.

MPP

MPP = Massively Parallel Processing
Nodi con risorse dedicate e comunicazione distribuita.

Le basi dati ad hoc possono essere ospitate anche in fogli di calcolo.

I tipi di struttura dei dati

Struttura e formato influenzano il metodo di analisi

STRUTTURATI

Dati organizzati in campi e record.

Esempi: tabelle e CSV (Comma-Separated Values).

SEMI-STRUTTURATI

Struttura flessibile e metadati.

JSON (JavaScript Object Notation), XML (eXtensible Markup Language), documenti con metadati.

NON STRUTTURATI

Non organizzati in campi tabellari.

Esempi: testo libero, immagini, video e audio.

Trattamento dei dati non strutturati

Esempio: sentiment analysis su commenti social

1
TESTO
Tweet e post

→

2
PREPROCESSING
Pulizia e normalizzazione

→

3
MATRICE
Document-term matrix

→

4
ANALISI
Tecniche statistiche e predittive

Document-term matrix: righe = documenti · colonne = parole · celle = frequenze

La provenienza dei dati

Le informazioni possono essere prodotte da persone o macchine

DATI GENERATI DALLE PERSONE

- Social network
- E-commerce
- Inserimento manuale nei sistemi

DATI GENERATI DALLE MACCHINE

- Sensori e strumenti di misura
- Log e sistemi informatici
- Lettori di codici a barre

Anche i dati automatici possono presentare anomalie: ad esempio un sensore guasto.

Attori aziendali e dati

Tre prospettive complementari sui dati aziendali

MANAGER

KPI (Key Performance Indicator) e analisi di sintesi.
Decisioni strategiche e tattiche.
Previsioni e scenari di business.

PERSONALE OPERATIVO

Dati granulari e tempestivi.
Gestione delle attività quotidiane.
ERP (Enterprise Resource Planning),
SCM (Supply Chain Management),
CRM (Customer Relationship Management).

DATA SCIENTIST

SQL (Structured Query Language), Python/R, Hadoop, Spark, NoSQL (Not Only SQL).
ML (Machine Learning), statistica e visualizzazione.
Qualità dei dati e dominio.

Archiviazione dei dati

Dai semplici file ai sistemi dedicati

FILE SYSTEM E ARCHIVI FLAT

CSV (Comma-Separated Values), JSON (JavaScript Object Notation), XML (eXtensible Markup Language), TXT (Text), Excel, Parquet.
File locali o cloud; scambio dati e archivi storici.

FOGLI DI CALCOLO

Excel e Google Sheets.
Analisi preliminari e personali.
Limiti in scalabilità, tracciabilità e sicurezza.

Archiviazione dei dati: database

Due grandi famiglie di sistemi di gestione

DATABASE RELAZIONALI (RDBMS)

RDBMS = Relational Database Management System.
Tabelle, righe, colonne; query SQL.
Esempi: PostgreSQL, MySQL, Oracle.
Uso: ERP, CRM e contabilità.

DATABASE NON RELAZIONALI (NoSQL)

NoSQL = Not Only SQL.
Modelli: documenti, key-value, grafi e colonne.
Esempi: MongoDB, Redis, Neo4j, Cassandra.
Uso: web, social, log e IoT.

Archiviazione dei dati: analisi

Data warehouse e data mart

DATA WAREHOUSE

Ottimizzato per analisi storiche e reporting.
Dati integrati e strutture favorevoli a interrogazioni complesse.
Esempi: BigQuery, Redshift, Snowflake, Teradata.

DATA MART

Sottoinsieme tematico dei dati analitici.
Concentrato su un reparto, come marketing o vendite.
Più mirato e semplice da consultare.

Archiviazione dei dati: Data Lake

Un repository flessibile per dati grezzi ed eterogenei

DATA LAKE

Conserva grandi volumi di dati nel formato originale.

Gestisce dati strutturati, semi-strutturati e non strutturati.

Supporta Data Science, Machine Learning e analisi avanzata.

Tecnologie: Hadoop, Azure Data Lake, AWS (Amazon Web Services).

RIFERIMENTO BIBLIOGRAFICO

Big Data Analytics

Capitolo 1 — A. Rezzani