BIG DATA ANALYTICS

COSA SONO I BIG DATA

Prof.ssa Romina Eramo

Università degli Studi di Teramo

Dipartimento di Scienze della Comunicazione

reramo@unite.it

Introduzione ai Big Data

Dall’espressione di moda a una capacità concreta di analisi.

Attorno al 2011

Il termine si diffonde come buzzword: promette nuove opportunità, spesso prima di vere applicazioni.

Oggi

Le organizzazioni raccolgono e analizzano grandi quantità di dati, strutturati e non strutturati.

Open source + cloud → costi più accessibili, maggiore granularità e profondità storica

Tipologie di analisi

Dalla comprensione del passato al supporto delle decisioni.

01

DESCRITTIVA

Che cosa è successo?

Misura i fenomeni osservati nel passato.

02

PREDITTIVA

Che cosa potrebbe accadere?

Anticipa eventi e comportamenti futuri.

03

PRESCRITTIVA

Quale azione conviene scegliere?

Suggerisce azioni e strategie ottimali.

Le analisi avanzate possono generare vantaggi competitivi.

Definizione di Big Data

Le 3V descrivono la complessità dei dati; contano anche i limiti delle tecnologie.

V **VOLUME**

Grandi quantità di dati

Da terabyte in su, a seconda del contesto.

V **VELOCITÀ**

Dati generati di continuo

Sensori e dispositivi IoT ad alta frequenza.

V **VARIETÀ**

Fonti e formati diversi

Dati strutturati, semi-strutturati e non.

DEFINIZIONE OPERATIVA

Big Data: dati che non si possono analizzare con tecnologie tradizionali, o per cui farlo è troppo oneroso.

Limiti delle tecnologie tradizionali

Perché una soluzione basata solo su database relazionali può non bastare.

SOLUZIONI TRADIZIONALI

RDBMS

Database relazionali e strumenti analitici progettati soprattutto per dati organizzati in tabelle.

01

DATI NON STRUTTURATI

Gestione meno naturale di formati eterogenei.

02

COSTI E VOLUMI

Costi elevati al crescere della quantità di dati.

03

DATI VELOCI

Difficoltà con flussi in tempo reale o ad alta frequenza.

Big Data = Dati + Tecnologie

Il termine indica sia una tipologia di dati sia i sistemi per gestirli.

DATI

Volume • Velocità • Varietà

Dati complessi e eterogenei

+

TECNOLOGIE

Acquisizione • Storage • Analisi

Hadoop, Spark, cloud e sistemi distribuiti

L’obiettivo: estrarre valore da dati difficili da trattare con i soli strumenti tradizionali.

Hadoop e Spark

Due tecnologie chiave, con ruoli differenti e complementari.

HADOOP

Framework open source per il calcolo distribuito su cluster di computer.

HDFS → file system distribuito
YARN → gestione risorse
MapReduce → elaborazione batch

GESTIONE DISTRIBUITA DEI DATI

SPARK

Motore di calcolo distribuito flessibile.

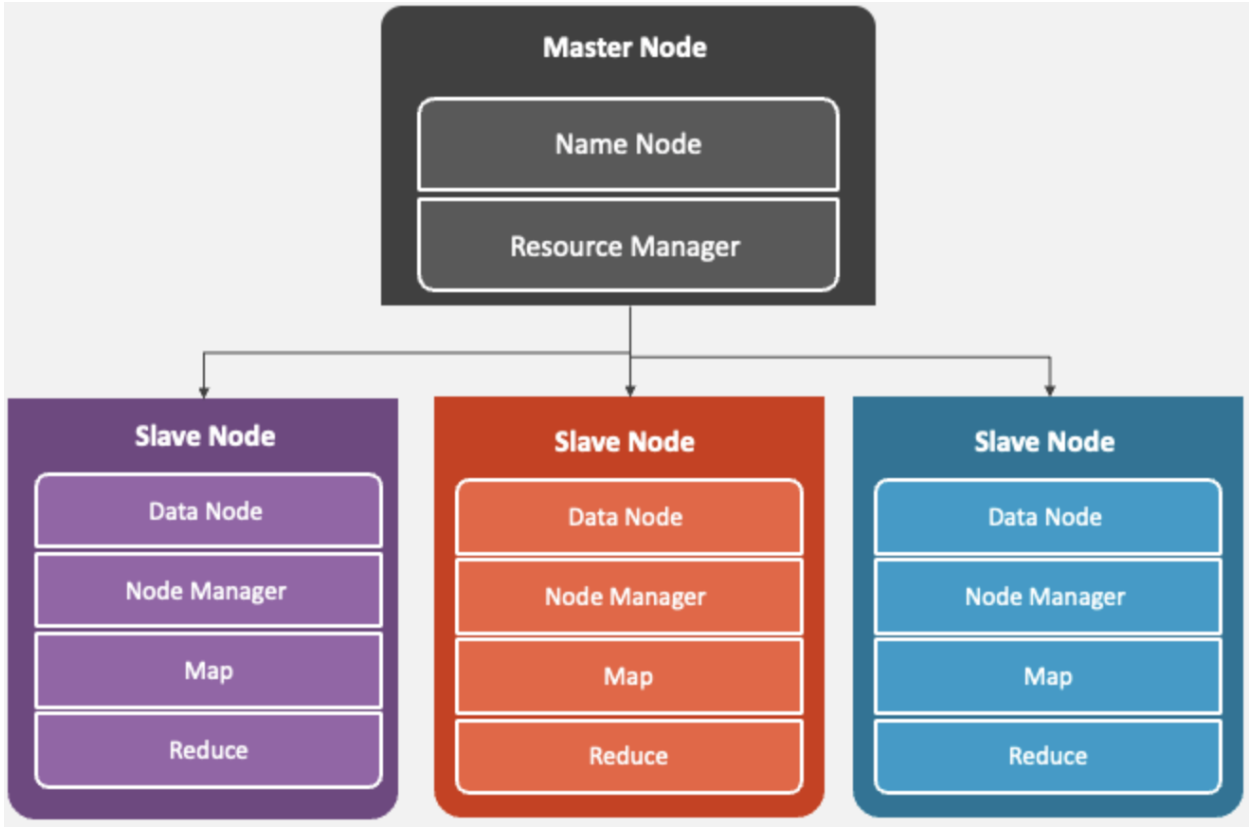
Supporta elaborazioni SQL, streaming e machine learning.

Può operare con Hadoop o indipendentemente.

ELABORAZIONE E ANALISI

Dentro un cluster Hadoop

Un nodo master coordina più nodi che memorizzano ed elaborano i dati.



ARCHITETTURA

Master

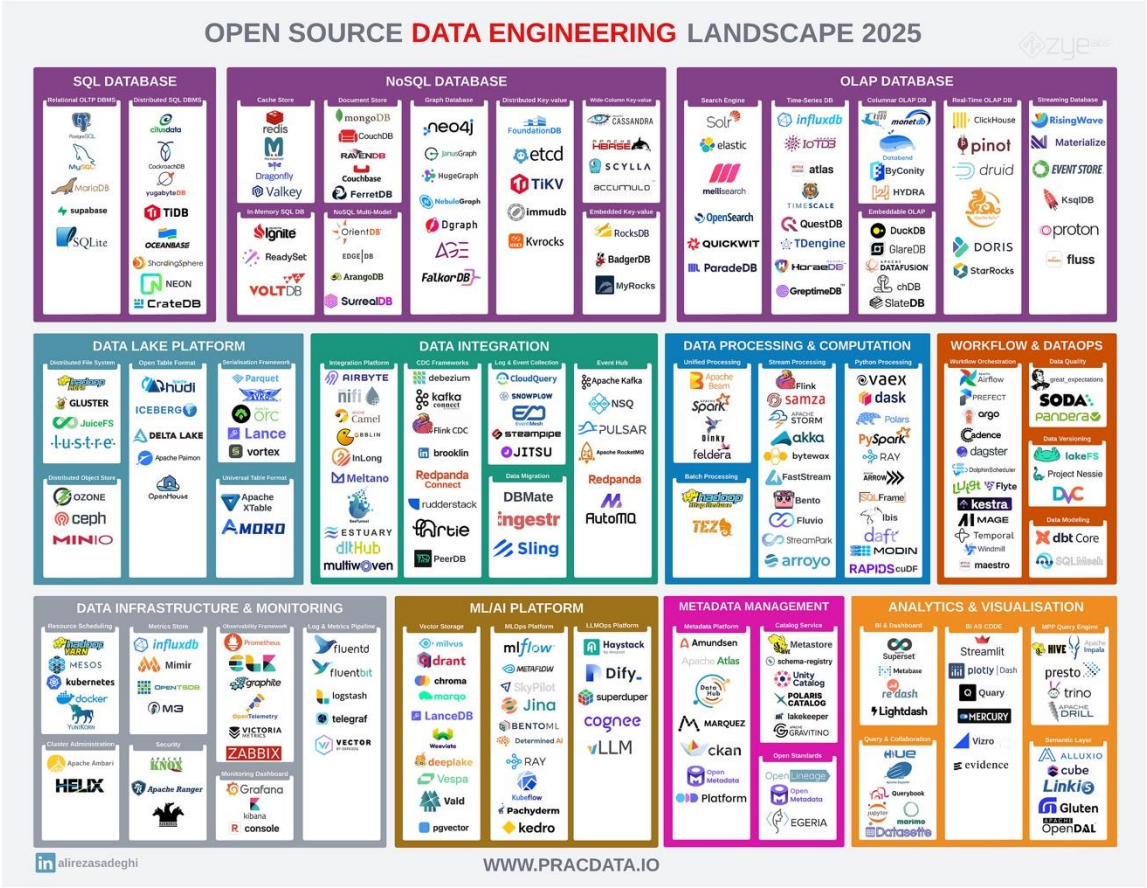
Coordinamento e gestione delle risorse.

Worker

Conservazione ed elaborazione distribuita.

Ecosistema open source: molte tecnologie

Mappa illustrativa delle soluzioni per archiviazione, integrazione, elaborazione e analytics (2025).



Tipologie di Big Data: esempi

I dati possono differire per volume, velocità e varietà.

FONTE / CASO	CARATTERISTICA	APPLICAZIONE
Sensori e DCS	Velocità e volume	Guasti e manutenzione predittiva
RFID	Velocità e volume	Percorso d'acquisto nella GDO
Transazioni finanziarie	Velocità e volume	Trading ad alta frequenza, previsioni
Strumenti scientifici	Velocità e volume	Pattern e simulazioni
Dati astronomici	Volume e varietà	Analisi astronomiche
Dati meteorologici	Volume	Previsioni e fenomeni estremi
Informazioni sanitarie	Volume e varietà	Monitoraggio delle malattie
Dati fiscali e bancari	Volume	Individuazione di possibili evasioni
Social network	Varietà (semi-strutt.)	Sentiment analysis, CRM, intelligence
Blog e forum	Varietà (semi-strutt.)	Sentiment analysis, intelligence
Log dei server web	Volume	Traffico web e navigazione
Log dei router	Volume e velocità	Analisi di rete e provider

Esempi non esaustivi

Altri esempi di Big Data

Alcune fonti non sono naturalmente riconducibili a una tabella.

SORVEGLIANZA

Volume, velocità e varietà.

Analisi di video e dati per sicurezza e controllo.

DOCUMENTI

Volume e assenza di struttura.

Analisi delle richieste di risarcimento per individuare frodi sospette.

DATI GEOGRAFICI

Volume e velocità.

Integrazione di dati GIS con altre fonti per diverse applicazioni.

La varietà dei dati condiziona strumenti e tecniche di analisi.

Tecnologie Big Data lungo il ciclo dei dati

Tre fasi fondamentali, spesso implementate da strumenti differenti.

01

ACQUISIZIONE

Raccogliere dati dalle fonti

data ingestion

02

STORAGE E ORGANIZZAZIONE

Conservare dati su larga scala

data storage

03

TRASFORMAZIONE E ANALISI

Preparare e interpretare i dati

data processing

Acquisizione dei dati

Modalità differenti per dati esistenti e per flussi in continuo.

DATABASE

RDBMS

Connettori, ETL (Extract, Transform, Load), Sqoop

Esempi: HDFS, HBase, NoSQL

FLUSSI

Dati in tempo reale

Strumenti di streaming

Esempi: Flume, Storm, Kafka

SERVIZI

Piattaforme esterne

Interfacce di accesso

ODBC e API (Application Programming Interface)

Le API permettono ai sistemi software di comunicare e recuperare dati da altre applicazioni.

Immagazzinamento e organizzazione

Hadoop e NoSQL offrono alternative per dati voluminosi e non strutturati.

HADOOP: COMPONENTI CHIAVE

HDFS	File system distribuito
YARN	Gestione delle risorse nel cluster
MapReduce / Tez	Elaborazione parallela

NOSQL

Database non relazionali

Adatti a modelli di dati flessibili e a scenari distribuiti.

Storage distribuito: Hadoop e NoSQL

Distinguiamo file system, motori di calcolo e database.

1 **HDFS**
Conserva dati eterogenei con ridondanza.

2 **MapReduce**
Suddivide un calcolo in sotto-attività parallele.

3 **HBase**
Supporta accessi tipici di un database sopra l'ecosistema Hadoop.

DATABASE NOSQL

Cassandra · MongoDB · Neo4j

Database non relazionali che supportano esigenze di organizzazione e scalabilità differenti.

HDFS non è un database: ogni livello risolve una necessità specifica.

Trasformazione e analisi dei dati

Strumenti che riducono la complessità delle elaborazioni distribuite.

MAPREDUCE

Elaborazione nativa

Potente, ma richiede sviluppo dedicato (tipicamente Java).

PIG

Trasformazione procedurale

Pig Latin semplifica le trasformazioni sui dati.

HIVE

Interrogazioni SQL-like

HiveQL supporta query, aggregazioni e analisi di grandi dataset.

Trasformazione e analisi restano fasi diverse, anche se condividono strumenti.

Strumenti di analisi avanzata

Dal machine learning statistico all’elaborazione distribuita.

MAHOUT

Machine learning

Classificazione, clustering e motori di raccomandazione.

SPARK

Analytics distribuita

Streaming, SQL, trasformazioni e librerie di machine learning.

R

Statistica e ML

Linguaggio open source integrabile con Hadoop tramite connettori.

COMBINABILI IN ARCHITETTURE BIG DATA

Casi applicativi

Quattro contesti in cui l'analisi di dati su larga scala produce valore.

BANKING

Clienti, transazioni, rischi

INDUSTRY 4.0

Impianti e manutenzione

INTERNET OF THINGS

Oggetti e sensori connessi

SMART CITY

Mobilità, servizi, ambiente

Il settore bancario

I Big Data permettono analisi più granulari su serie storiche più estese.

LA SFIDA

Milioni di operazioni

Le sintesi mensili possono nascondere dettagli importanti: servono dati storici e transazioni granulari.

HADOOP

Conservazione e analisi della storia delle transazioni

SPARK

Calcolo di modelli predittivi su grandi dataset

ANALISI DEL TESTO

Valorizzazione di campi descrittivi (es. causali dei bonifici)

ESEMPI: CHURN · CLUSTERING · TARGETING

Churn, clustering e targeting

Tre applicazioni per comprendere e raggiungere meglio i clienti.

CHURN PREDICTION

Chi potrebbe abbandonare?

Intervenire in anticipo per fidelizzare.

CLUSTERING

Quali clienti si somigliano?

Segmentare la clientela in gruppi omogenei.

CAMPAIGN TARGETING

Chi è più ricettivo?

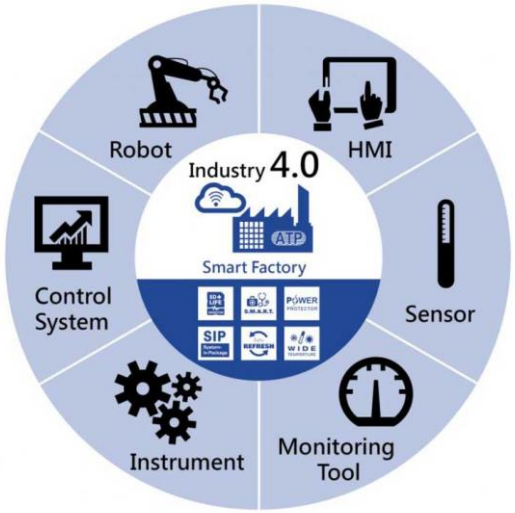
Rendere le campagne più efficaci e mirate.

Industry 4.0

Produzione industriale automatizzata, connessa e orientata ai dati.

FATTORI ABILITANTI

- Big Data
- Internet of Things
- Cloud Computing
- Machine Learning



Sensori e stream analytics → Monitoraggio → Previsione dei guasti → Manutenzione programmata

Internet of Things (IoT)

Oggetti connessi che producono e inviano dati sul proprio utilizzo.

CHE COS'È

Sensori + connettività

I dispositivi diventano fonti di dati: automobili, macchine industriali, oggetti di uso quotidiano.

OGGETTI → DATI → ANALISI

VEICOLI

Scatole nere, analisi stile di guida

AUTO CONNESSE

Manutenzione, percorsi e infotainment

DOMOTICA

Controllo e automazione domestica

AGRICOLTURA / RFID

Sensori ambientali e flussi di visitatori

Le smart city

I dati sostengono decisioni e servizi per la città.

OBIETTIVI

- Efficienza energetica
- Mobilità urbana
- Sicurezza
- Riduzione dell’inquinamento



Fonti: sensori, dispositivi IoT e social network.

Caso studio: Yinchuan (Cina)

Esempio riportato nel materiale didattico: dati per organizzare servizi urbani.





RIFIUTI

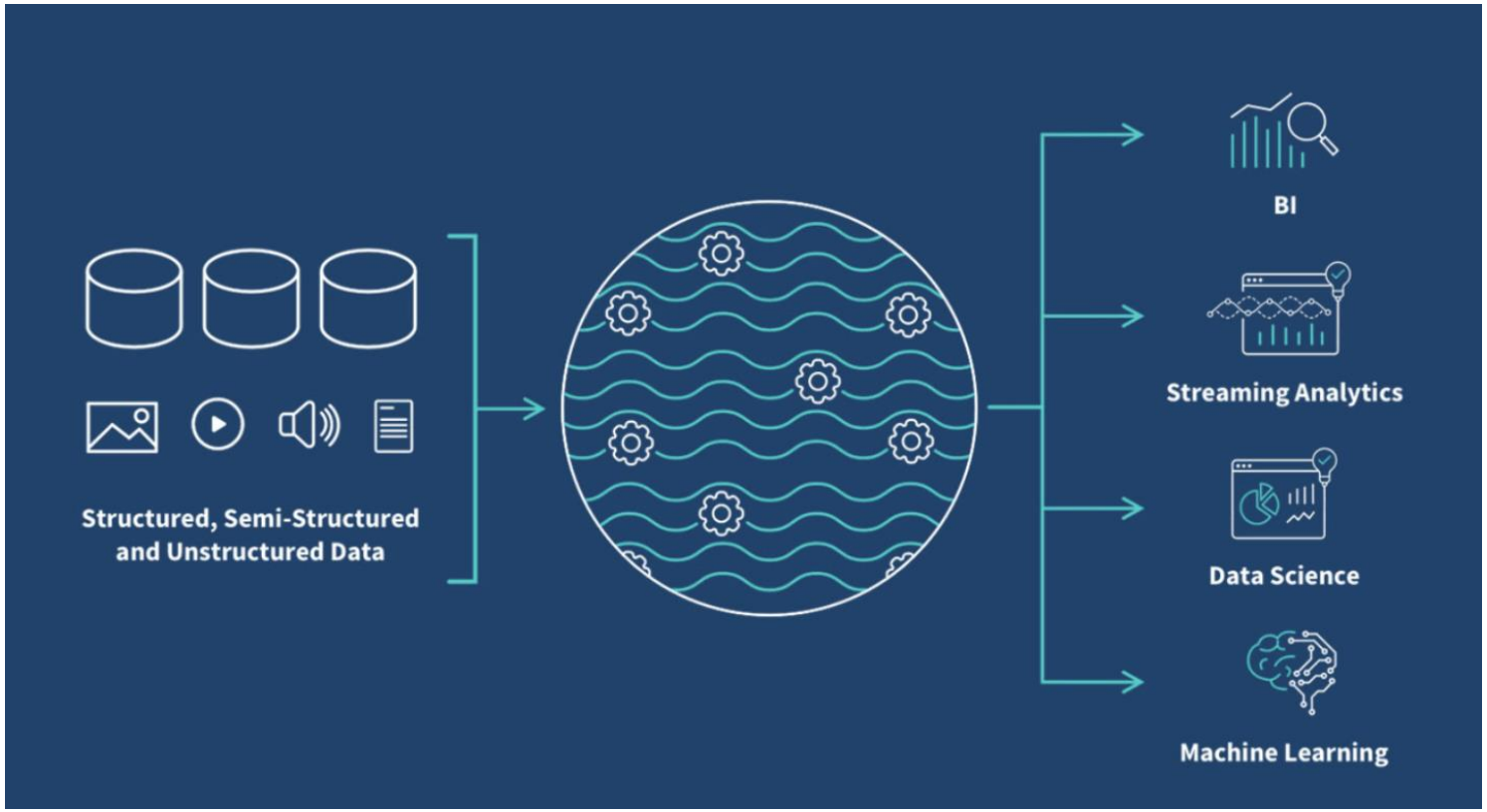
Contenitori intelligenti per ottimizzare la raccolta.

TRAFFICO

RFID e telecamere per monitorare i flussi; semafori adattivi.

L’architettura Data Lake

Un punto di raccolta per dati grezzi provenienti da molteplici fonti.



Nel Data Lake avvengono ingestion, storage, trasformazione e analisi dei dati.

Caratteristiche del Data Lake

Flessibilità e scala richiedono un’efficace data governance.

01 FONTI ETEROGENEE

Centralizza dati provenienti da sistemi diversi.

02 FORMATO ORIGINALE

Conserva dati grezzi; schema on read.

03 STORIA E DETTAGLIO

Mantiene grandi volumi e serie storiche.

04 USI DIFFERENTI

Supporta data scientist e altre figure.

05 GOVERNANCE

Metadati, sicurezza, controllo accessi e ricerca.

Lambda Architecture

Un’architettura che combina dati storici e aggiornamenti in tempo reale.

BATCH LAYER

Conserva tutti i dati ed esegue elaborazioni periodiche.

SPEED LAYER

Elabora i dati più recenti con aggiornamenti rapidi.

SERVING LAYER

Ottimizza viste e accesso ai dati per le interrogazioni.

OBIETTIVI · Scalabilità · Tolleranza ai guasti · Bassa latenza

Data Quality e Data Management Platform

La qualità del dato è necessaria per prendere decisioni attendibili.

DATA QUALITY

Gestione di metadati e verifiche.

Identificare e validare dati incompleti, incoerenti o poco affidabili, per stabilire quando possano essere utilizzati.

DATA MANAGEMENT PLATFORM

Integra dati aziendali e di terze parti.

Supporta analisi predittive e attività di pubblicità, CRM, marketing e targeting.

Data Lake e Data Warehouse

Architetture complementari nel sistema informativo aziendale.

DATA LAKE

Dati grezzi, molti formati, analisi complesse e archiviazione storica.

DATA WAREHOUSE

Dati strutturati e sintesi per reporting e business intelligence.

Il Lake può alimentare il Warehouse con risultati sintetici, o estenderne lo storage.

Pro e contro del Data Lake

La flessibilità è preziosa, ma non elimina costi e responsabilità.

PUNTI DI FORZA

- Grande flessibilità
- Schema on read
- Molti dati, anche non strutturati
- Supporto ad analisi predittive

CRITICITÀ

- Interrogazioni rapide non sempre ideali
- Competenze tecniche dedicate
- Costi di gestione e presidio
- Necessità di governance continua

Senza governance, il Data Lake può trasformarsi in un «data swamp».

Riferimento bibliografico

Materiale di base per l’introduzione alle tecnologie Big Data.

TESTO DI RIFERIMENTO

A. Rezzani

Capitolo 2 – Big Data Analytics