UNIVERSITÀ DEGLI STUDI DI TERAMO

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

» Il fenomeno Big Data nasce attorno al 2011

» Inizialmente una buzzword, oggi è una realtà concreta

» Obiettivo: estrarre valore da grandi volumi di dati, strutturati e non

» **Tecnologie e diffusione**

- Tecnologie open source e cloud hanno ridotto i costi di storage e licenze
- Analisi ad alta granularità e profondità storica

Tipologie di analisi

» **Descrittiva**: guarda al passato e misura cosa è accaduto

» **Predittiva**: anticipa eventi futuri

» **Prescrittiva**: suggerisce azioni ottimali

» Le analisi avanzate generano vantaggi competitivi

Definizione di Big Data

» I Big Data si identificano secondo le **3V**:
  - **Volume**: enormi quantità di dati (terabyte e oltre)
  - **Velocità**: dati generati ad alta frequenza (es. sensori, IoT)
  - **Varietà**: dati provenienti da fonti diverse, in formati strutturati e non

» **Oltre le 3V: definizione moderna**
  - I Big Data sono:
    - » Dati non analizzabili con tecnologie tradizionali
    - » Oppure dati per cui le tecnologie tradizionali sono troppo costose
  - La definizione include quindi un **criterio economico e tecnologico**

Limiti delle tecnologie tradizionali

» Tecnologie tradizionali:
  - Database relazionali (RDBMS)
  - Strumenti analitici basati su dati strutturati

» Limiti:
  - Difficoltà a gestire dati non strutturati
  - Elevati costi con grandi volumi
  - Prestazioni ridotte su dati in tempo reale o ad alta velocità

Big Data = Dati + Tecnologie

» Il termine Big Data comprende:
  - **Dati complessi** per volume, velocità, varietà
  - **Tecnologie** per ingestione, storage e analisi

» Soluzioni: Hadoop, Spark, sistemi distribuiti, cloud scalabile

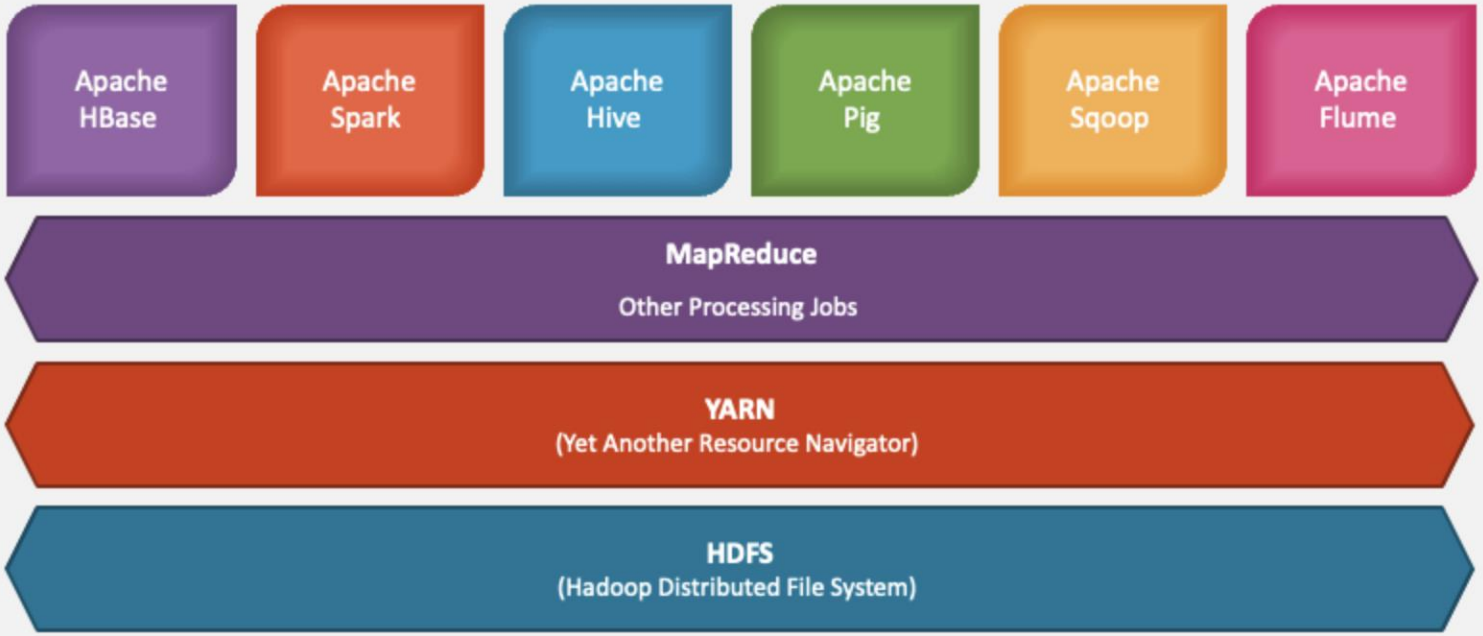
Hadoop e Spark: tecnologie chiave per i Big Data

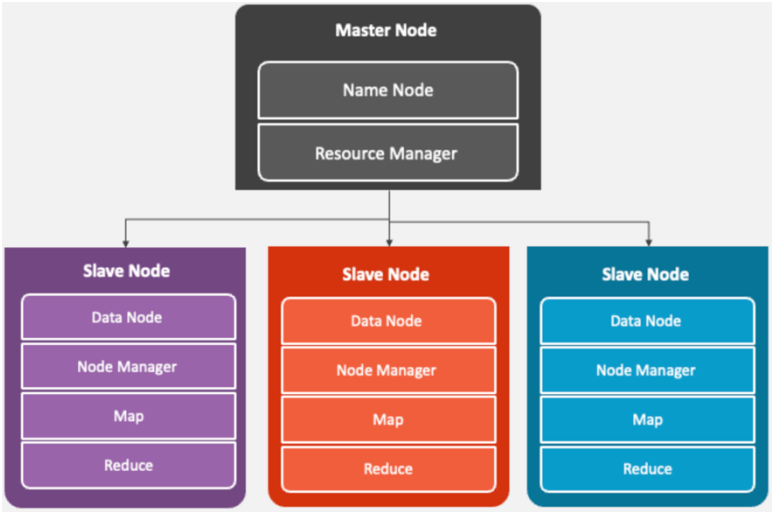
» Hadoop
  - Sistema open source per il **calcolo distribuito**
  - Gestisce grandi volumi di dati su cluster di computer
  - Componenti principali:
  - **HDFS** (file system distribuito)
  - **YARN** (gestione risorse)
  - **MapReduce** (elaborazione batch)

» Spark
  - Motore di calcolo distribuito **più veloce e flessibile** di MapReduce
  - Supporta **analisi in tempo reale**, **streaming**, **SQL**, **machine learning**
  - Può lavorare **insieme a Hadoop** o su sistemi indipendenti
  - Usato per analisi complesse su grandi volumi di dati

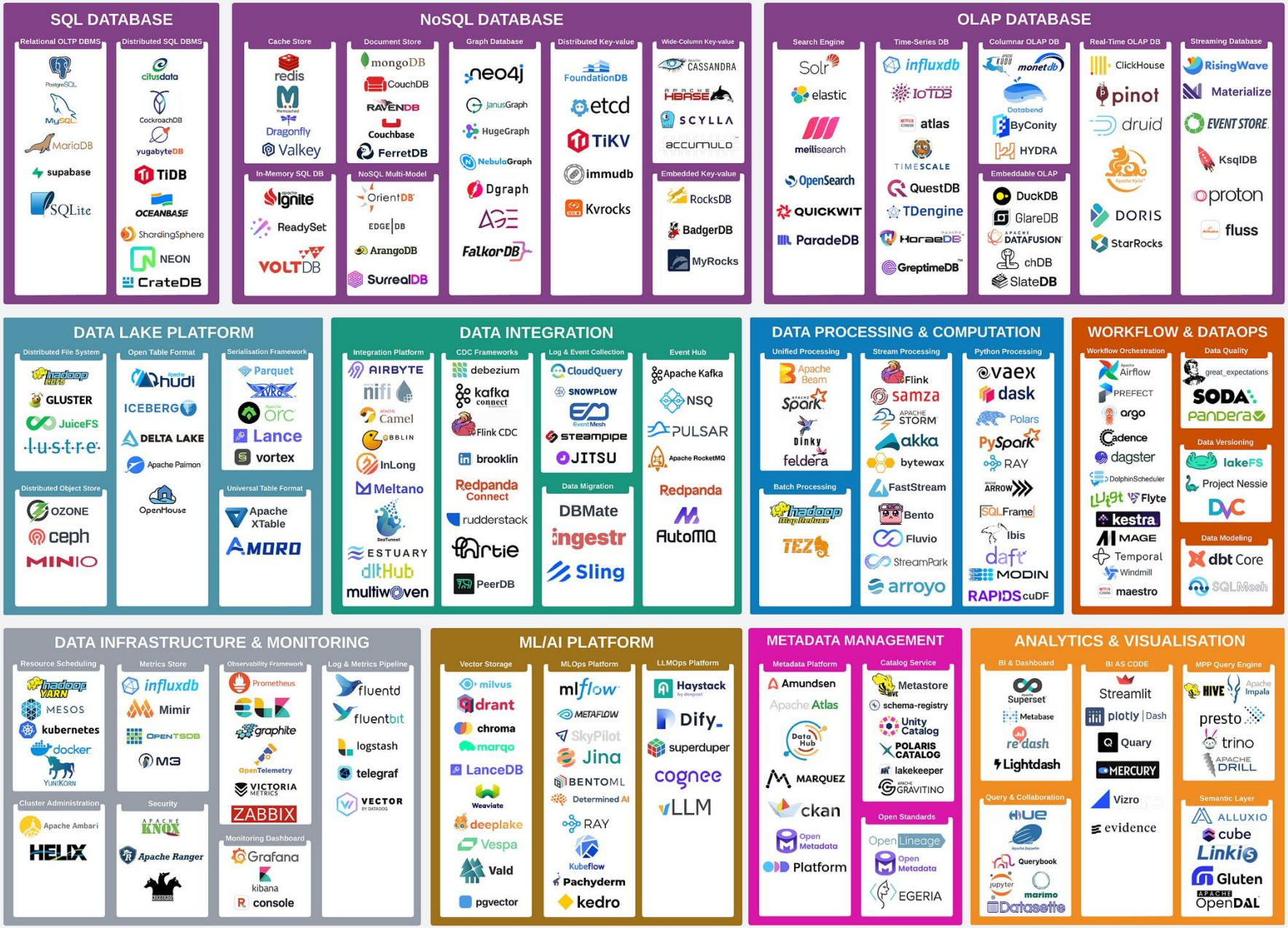
HADOOP ARCHITECTURE

Hadoop Architecture and Ecosystem





OPEN SOURCE DATA ENGINEERING LANDSCAPE 2025



alirezasadeghi

Tipologie di Big Data (esempi)

Caso	Caratteristiche	Esempi di utilizzo
Sensori e DCS	Velocità e volume	Analisi guasti, manutenzione predittiva
RFID	Velocità e volume	Percorso d'acquisto in GDO
Quotazioni e transazioni finanziarie	Velocità e volume	High frequency trading, analisi previsionali
Dati da strumenti scientifici	Velocità e volume	Riconoscimento pattern, simulazioni
Dati astronomici	Volume e varietà	Analisi astronomica avanzata
Dati metereologici	Volume	Previsioni meteo, eventi estremi
Informazioni sanitarie	Volume e varietà	Monitoraggio malattie, enti sanitari
Dati fiscali, bancari e patrimoniali	Volume	Identificazione evasione fiscale
Social Network	Varietà (semi-strutturati)	Sentiment analysis, CRM, intelligence
Blog, Forum	Varietà (semi-strutturati)	Sentiment analysis, servizi di intelligence
Web server log	Volume	Analisi traffico web, navigazione utenti
Log router	Volume e velocità	Analisi rete, provider

…

Tipologie di Big Data (esempi)

Caso	Caratteristiche	Esempi di utilizzo
Dati da sistemi di sorveglianza	Volume, velocità, varietà	Sicurezza pubblica, videosorveglianza, servizi di intelligence
Documenti	Volume e assenza di struttura	Fraud detection: analisi delle richieste di risarcimento per identificare potenziali frodi e sottoporle a controlli approfonditi.
Dati geografici	Volume, velocità	I dati provenientida sistemi GIS possono essere utilizzati assieme ad altri per scopi diversi

Tecnologie Big Data in breve

» Le tecnologie Big Data possono essere suddivise in base al ciclo di vita dei dati:

- **Acquisizione** (data ingestion)
- **Immagazzinamento e organizzazione**
- **Trasformazione e analisi**

Acquisizione

» L'acquisizione varia in base alla fonte e al formato dei dati

» **Dati da RDBMS**: tramite strumenti come **Sqoop** (Hadoop) o **ETL tradizionali** con connettori per HDFS, HBase, NoSQL

» **Dati in tempo reale**: gestiti con strumenti di **data streaming** come **Flume, Storm, Kafka**

» Le connessioni possono avvenire tramite **ODBC** o **API**

» **API** (es. **Twitter API**, **Facebook Graph API**, **Yahoo YQL**) permettono l'accesso a contenuti pubblici o in tempo reale

Immagazzinamento e organizzazione

» Le tecnologie tradizionali (es. RDBMS) faticano con grandi volumi e dati non strutturati

» Le principali soluzioni: **Hadoop** e **database NoSQL**

» **Hadoop** è un sistema open source per il calcolo distribuito, affidabile e scalabile

» Componenti chiave:

- **HDFS** – file system distribuito
- **YARN** – gestione delle risorse nel cluster
- **MapReduce / Tez** – elaborazione distribuita in parallelo

Immagazzinamento e organizzazione (2)

» HDFS garantisce la **ridondanza dei dati** e supporta ogni tipo di formato

» **MapReduce** divide i task in sotto-task elaborati in parallelo

» L'**ecosistema Hadoop** include Spark, HBase, Accumulo e altri strumenti Apache

» HDFS non è un database: per accesso rapido ai dati si usano strumenti come **HBase**

» I **NoSQL** (Cassandra, MongoDB, Neo4j…) sono database non relazionali, distribuiti e scalabili

» Il volume si focalizzerà su Hadoop e componenti integrati (es. HBase, Accumulo)

Trasformazione e analisi dei dati

» Le fasi di trasformazione e analisi sono concettualmente distinte, ma condividono molti strumenti
  - **MapReduce**: strumento nativo di Hadoop, potente ma complesso (richiede Java)
  - **Pig**: linguaggio procedurale (Pig Latin), semplifica le trasformazioni su Hadoop
  - **Hive**: sistema di data warehousing su Hadoop, consente analisi SQL-like con HiveQL
  - Hive e Pig semplificano l'interazione con l'ecosistema Hadoop per utenti non esperti

Trasformazione e analisi dei dati

» Strumenti avanzati per l'analisi nei Big Data
  - **Mahout**: piattaforma per machine learning (clustering, classificazione, recommendation engine)
  - **Spark**: motore distribuito avanzato con supporto per:
  - Data ingestion (anche streaming)
  - Trasformazioni e analisi SQL
  - Machine learning integrato
  - Spark + Hadoop: accoppiata sempre più usata per efficienza e scalabilità
  - **R**: linguaggio open source per statistica e ML, integrabile con Hadoop tramite connettori

Casi interessanti

» Il settore banking

» Industry 4.0

» Internet of Things

» Le smart city

Il settore banking

» Diverse grandi banche italiane adottano sistemi Big Data come **Hadoop** e **Spark**

» Gli strumenti non sempre sono usati per analisi sofisticate, ma per superare i **limiti tecnici** dei volumi elevati

» Si estendono analisi già in uso (es. saldi mensili) verso dati **granulari e storici**

» Hadoop consente **storicizzazione efficiente** e accesso con latenze contenute

» Spark permette **modelli predittivi** su grandi dataset in tempi ragionevoli

» Esempi: **churn prediction, clustering, campaign targeting ***

» Si usano anche tecniche di analisi del testo (es. causali bonifici) per **personalizzazione e prevenzione churn**

* Tecniche di analisi: Churn, Clustering, Campaign Targeting

» Churn Prediction
- Previsione dei clienti che potrebbero abbandonare l'azienda

→ Obiettivo: intervenire in tempo per fidelizzare

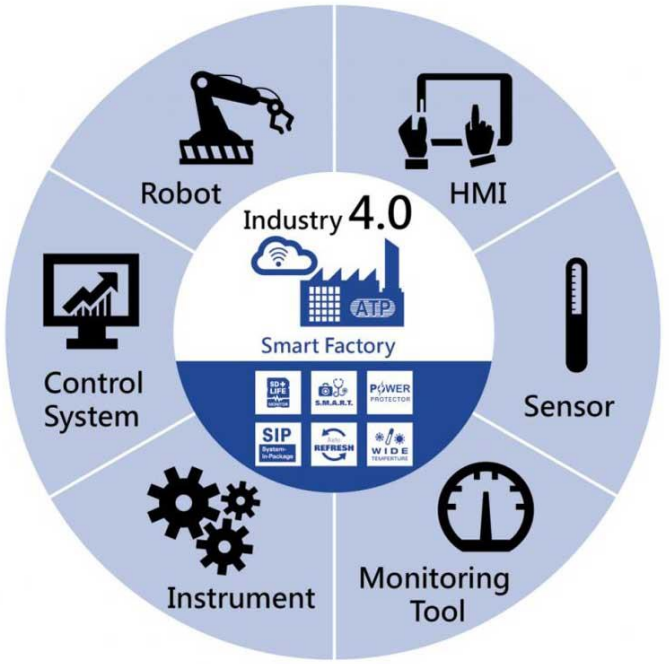
» Clustering
- Raggruppamento di clienti con caratteristiche o comportamenti simili

→ Obiettivo: segmentare la clientela per strategie mirate

» Campaign Targeting
- Indirizzare campagne verso i clienti più ricettivi

→ Obiettivo: aumentare efficacia e ritorno delle campagne di marketing

Industry 4.0



» Industry 4.0 è un percorso verso la **produzione industriale automatizzata e interconnessa**

» Non è solo una buzzword: richiede **competenze, tempo e investimenti**

» Impatti non ancora del tutto quantificabili: **costi, produttività, occupazione**

» **Fattori abilitanti**:
- Big Data
- Internet of Things (IoT)
- Cloud Computing
- Advanced Analytics / Machine Learning

» **Applicazioni: monitoraggio e manutenzione predittiva**
- **Analisi quasi in tempo reale** dei dati di impianto (es. da sensori)
- Uso di strumenti di **stream analytics** per acquisizione ed elaborazione simultanea
- Monitoraggio visivo e continuo della produzione
- Integrazione con **tecniche predittive** → proactive maintenance
- Si creano modelli che **anticipano guasti**, consentendo **fermi programmati** e riduzione dei costi

Internet of Things

» Oggetti dotati di **sensori e connettività** che producono e inviano dati

» Rientrano anche le macchine industriali (Industry 4.0) e dispositivi consumer

» **Esempi applicativi**:

- **Scatole nere** nei veicoli → monitoraggio stile di guida, sconti RCA, controllo sinistri
- **Auto connesse** → dati da sensori per manutenzione predittiva, infotainment, navigazione ottimizzata
- **Domotica** → automazione e controllo degli ambienti domestici
- **Agricoltura smart** → monitoraggio climatico con sensori (umidità, temperatura, pioggia…)
- **RFID nei parchi a tema** → tracciamento flussi visitatori, ottimizzazione personale e rifornimenti

Le smart city

» Le smart city sfruttano **tecnologie digitali e dati** per migliorare:
  - Efficienza energetica
  - Mobilità urbana
  - Sicurezza
  - Riduzione dell'inquinamento

» I dati provengono da **sensori**, **dispositivi IoT** e **social network**

» Richiesto l'uso di tecnologie capaci di gestire **grandi volumi di dati** (anche non strutturati)

Esempio smart city: Yinchuan (Cina)

» **Contenitori intelligenti** per rifiuti → raccolta ottimizzata

» Gestione traffico con:

» **Etichette RFID** su oltre metà dei 500.000 veicoli

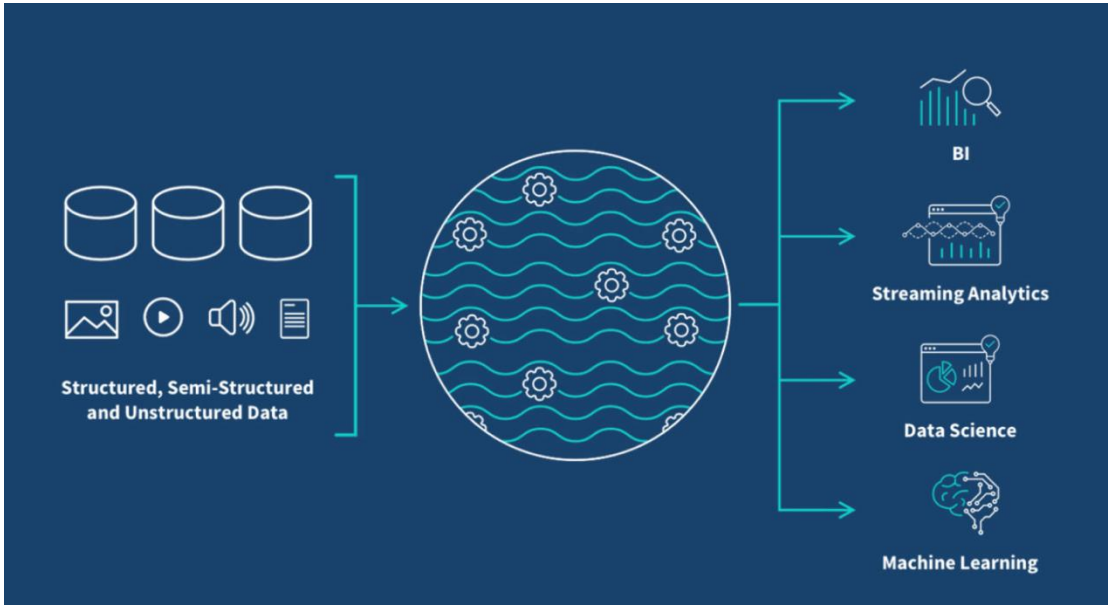
» **Telecamere + RFID** per monitoraggio e previsione flussi

» **Semafori adattivi** regolati dai dati in tempo reale

L’architettura Data Lake

» Cos’è un Data Lake

- Architettura per raccogliere, gestire e analizzare Big Data
- Supporta tutte le fasi: ingestion, trasformazione, analisi, storage
- Raccoglie dati grezzi anche non strutturati da molteplici fonti



Caratteristiche del Data Lake

» Centralizza dati da fonti eterogenee
» Salva dati nel formato originale (schema on read)
» Conserva grandi volumi di dati e profondità storica
» Accessibile a data scientist e altri utenti
» Richiede governance (metadati, sicurezza, ricerca)

Lambda Architecture

» Progettata per gestire diversi ritmi di arrivo e consumo dei dati

» Composta da tre livelli:

- **Batch layer**: conservazione ed elaborazione periodica
- **Speed layer**: aggiornamento in tempo reale
- **Serving layer**: ottimizza l'accesso ai dati

Data Quality & DMP

» La qualità dei dati va gestita con metadati e validazioni

» **DMP (Data Management Platform)**:

- Aggrega dati interni e da terze parti
- Supporta analisi predittive
- Usata per pubblicità, CRM, marketing, targeting

Integrazione con il sistema aziendale

» Il data lake **non sostituisce** il data warehouse

» Può agire come staging permanente, motore analitico o storage esteso

» Ideale per analisi complesse e dati non strutturati

» Output ridotto può alimentare il DWH

Pro e Contro del Data Lake

» ✅ Pro: flessibilità, schema on read, analisi predittiva, grandi volumi

» ❌ Contro: non adatto per interrogazioni interattive rapide

» Richiede skill specifici e governance

» Senza controllo rischia di diventare un **data swamp**

Capitolo 2 - Big Data Anlytics, A. Rezzani