Introduzione ai Big Data

Programmazione

I Big Data rappresentano un insieme di tecnologie, metodologie e processi progettati per gestire volumi di dati così elevati, eterogenei e dinamici da non poter essere trattati con i tradizionali sistemi di gestione dei dati.

Le caratteristiche principali vengono riassunte nel modello delle 3V (Volume, Velocità, Varietà), successivamente esteso con altre dimensioni come Veridicità e Valore.

Questo paradigma nasce dall'esigenza di analizzare dati provenienti da fonti diversificate sensori IoT, log di sistema, social network, transazioni finanziarie e trasformarli in informazioni strategiche utili per il supporto decisionale.

Dal punto di vista tecnico, l’ecosistema Big Data si fonda su architetture distribuite e scalabili, come Hadoop Distributed File System (HDFS) e framework di calcolo come Apache Spark e MapReduce.

Questi sistemi permettono di suddividere grandi dataset in blocchi, replicarli su cluster di nodi e processarli in parallelo, riducendo i tempi di elaborazione e aumentando la tolleranza ai guasti.

Inoltre, la scalabilità orizzontale consente di aggiungere risorse hardware al sistema senza dover ricorrere a costosi server centralizzati.

L’analisi dei Big Data richiede anche strumenti avanzati per l’elaborazione in tempo reale (streaming analytics), la gestione di dati non strutturati (documenti, immagini, video), e l’integrazione con modelli di Machine Learning e Intelligenza Artificiale.

Database NoSQL come MongoDB, Cassandra o HBase si affiancano ai sistemi relazionali per garantire flessibilità nello storage, mentre piattaforme di orchestrazione e cloud computing rendono possibile la gestione dinamica delle risorse.

In questo modo i Big Data non sono solo una questione di grande quantità di dati, ma soprattutto di capacità di estrarre valore da essi attraverso pipeline efficienti di acquisizione, elaborazione e visualizzazione.

Nei suoi albori, l’idea di raccogliere e analizzare grandi quantità di dati affonda le radici già nel XVII secolo: nel 1663, John Graunt, un mercante londinese, elaborò per primo statistiche sulla mortalità legate alla peste, gettando le basi del data analysis moderno.

Nei decenni successivi, lo sviluppo dell’informatica a partire dagli anni '60 con i primi sistemi IBM in grado di gestire grandi volumi di dati, passando per gli anni ’70 e ’80 con l’evoluzione dei database e dei software di elaborazione ha preparato il terreno per ciò che oggi definiamo Big Data.

l termine “Big Data”, nella sua accezione moderna, compare ufficialmente nella metà degli anni '90 ed è generalmente attribuito a John R. Mashey, chief scientist presso Silicon Graphics, che lo utilizzava nei suoi interventi tecnici per esempio in un talk del 1998 intitolato “Big Data and the Next Wave of Infrastress”

Nel 2001, Gartner coniò la celebre definizione delle “3V” Volume, Velocità e Varietà che ha contribuito a chiarire e diffondere il concetto nel mondo professionale e accademico

A partire dal secondo decennio del XXI secolo, l’esplosione di dati digitali grazie a internet, i social media, i dispositivi IoT e il cloud computing ha trasformato Big Data da termine tecnico a fenomeno di massa, dando così inizio all’era dei zettabyte

Caratteristiche principali (le 5 V "standard")

- **Volume**

Rappresenta la quantità immensa di dati generati e archiviati, spesso misurata in terabyte, petabyte o oltre.

La gestione di tali volumi richiede tecnologie di archiviazione distribuita e sistemi scalabili orizzontalmente.

- **Velocità (Velocity)**

Indica la rapidità con cui i dati vengono prodotti, trasmessi e processati, talvolta in tempo reale.

Questo comporta sfide legate alla latenza e alla capacità di analisi immediata, richieste per supportare decisioni tempestive.

- **Varietà (Variety)**

I dati provengono da fonti e formati molto diversi strutturati, semi-strutturati (come JSON, XML) o non strutturati (testi, immagini, video).

Richiede sistemi flessibili come NoSQL, data lake e tecnologie in grado di normalizzare dati disparati.

- **Veridicità (Veracity)**

Riguarda l’affidabilità, qualità e accuratezza dei dati.

Poiché i Big Data spesso provengono da fonti eterogenee e rumorose (ad esempio social media), è cruciale valutare l’incertezza, la presenza di errori o bias, e gestire la "qualità intrinseca" del dato.

- **Valore (Value)**

È l’utilità effettiva che si riesce a estrarre dai dati. I progetti di Big Data richiedono investimenti significativi: la vera sfida è trasformare dati grezzi in informazioni e insight utilizzabili, in grado di generare vantaggio competitivo o efficienza organizzativa.

Caratteristiche aggiuntive (ulteriori V)

Nel tempo, al modello delle 5 V se ne sono aggiunte di nuove per catturare meglio la complessità dei dati:

- **Variabilità (Variability)**

Rappresenta i cambiamenti nei formati o nei flussi dei dati e l'ambiguità semantica che può emergere (ad esempio durante eventi o trend passeggeri).

- **Visualizzazione (Visualization)**

Si riferisce alla capacità di rappresentare i Big Data in forma visiva — grafici, dashboard, mappe — per facilitarne l’interpretazione e l’analisi da parte degli stakeholder.

- ***Venue***

Indica la molteplicità dei sistemi o piattaforme dove i dati vengono archiviati e processati, come data lake, data warehouse, cloud e ambienti IoT.

- ***Vaghezza (Vagueness)***

Vocabolario evidenzia la necessità di termini e semantiche condivisi per interpretare i dati; Vaghezza delinea l'incertezza intrinseca nei dataset, spesso imperfetti o incompleti.

- ***Tabella riepilogativa***

Caratteristica	Significato tecnico
Volume	Enormi quantità di dati da archiviare e gestire a scala distribuita
Velocità	Rapidità di produzione e processazione dati, anche in tempo reale
Varietà	Diversità di formati e strutture dei dati da integrare
Veridicità	Accuratezza, affidabilità, qualità intrinseca dei dati
Valore	Capacità di ottenere insight utili e utilizzo concreto dei dati
Variabilità	Fluttuazioni nei formati/flussi dei dati nel tempo
Visualizzazione	Rappresentazione visiva per facilitare analisi e comprensione
Venue	Diversità di ambienti e piattaforme per gestione/archiviazione

Uno dei principali problemi legati ai Big Data riguarda la gestione della complessità intrinseca dei dati, che spesso risultano eterogenei, incompleti o rumorosi, rendendo difficile garantirne qualità e affidabilità.

A questo si aggiungono le sfide di scalabilità e performance, poiché i sistemi devono essere in grado di elaborare grandi volumi di dati in tempi ridotti, anche in scenari di streaming in tempo reale.

Altri ostacoli rilevanti sono legati alla sicurezza e alla privacy, vista la natura sensibile di molti dataset (ad esempio dati sanitari o finanziari), e alla conformità normativa, che impone vincoli stringenti sull'uso e la conservazione delle informazioni.

Infine, un problema ricorrente è la mancanza di competenze specializzate: senza figure in grado di progettare pipeline dati efficienti e interpretare correttamente i risultati, i Big Data rischiano di rimanere un grande investimento infrastrutturale privo di reale valore strategico.


Buon Davante a tutti